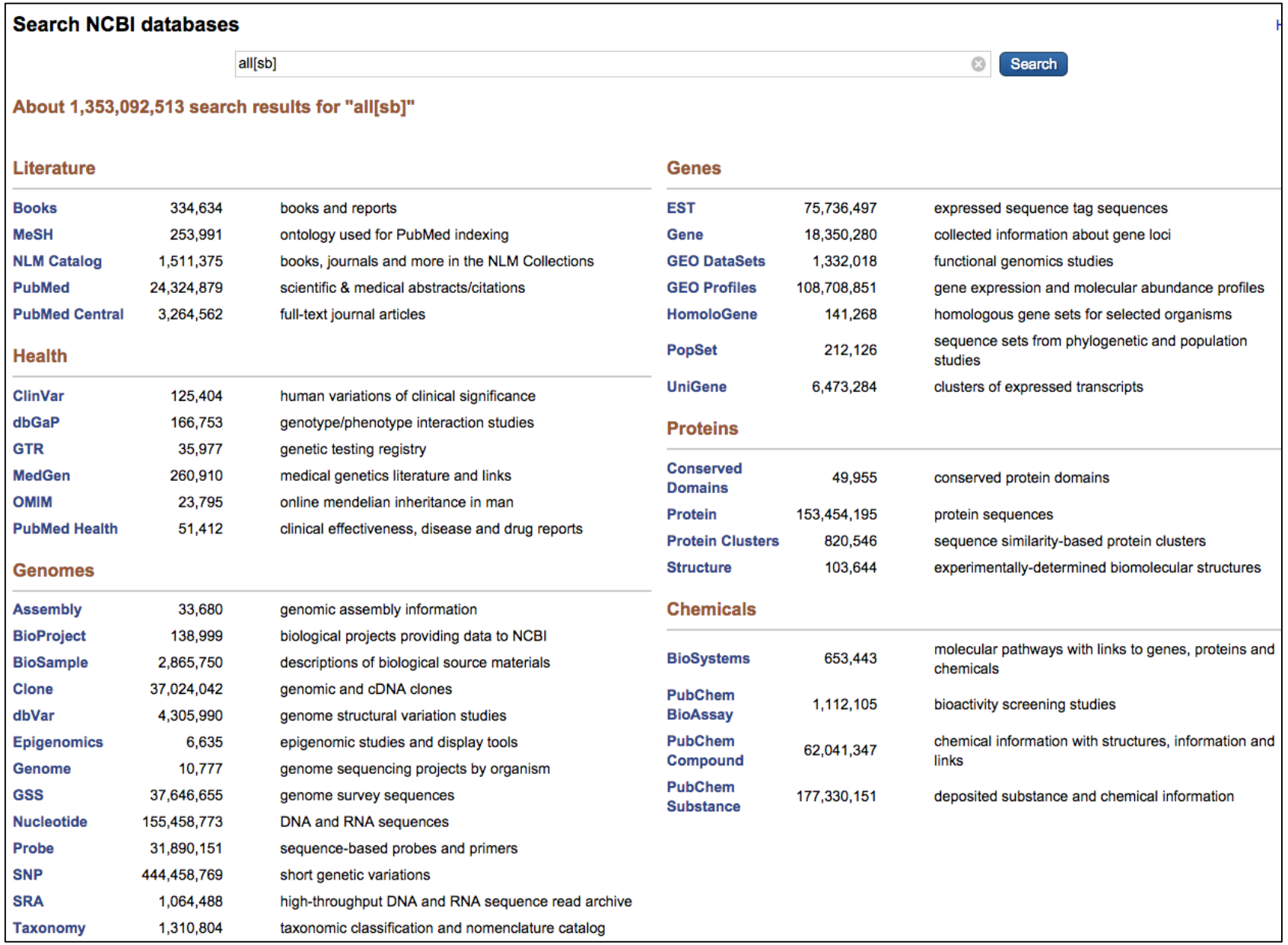
NCBI Entrez Utilities (EUtilities) and Direct(EDirect): Genomic Data Retrieval

Credit to 

Entrez is NCBI's primary text search and retrieval system



Search NCBI databases

all[sb] Search

About 1,353,092,513 search results for "all[sb]"

Literature

Books	334,634	books and reports
MeSH	253,991	ontology used for PubMed indexing
NLM Catalog	1,511,375	books, journals and more in the NLM Collections
PubMed	24,324,879	scientific & medical abstracts/citations
PubMed Central	3,264,562	full-text journal articles

Health

ClinVar	125,404	human variations of clinical significance
dbGaP	166,753	genotype/phenotype interaction studies
GTR	35,977	genetic testing registry
MedGen	260,910	medical genetics literature and links
OMIM	23,795	online mendelian inheritance in man
PubMed Health	51,412	clinical effectiveness, disease and drug reports

Genomes

Assembly	33,680	genomic assembly information
BioProject	138,999	biological projects providing data to NCBI
BioSample	2,865,750	descriptions of biological source materials
Clone	37,024,042	genomic and cDNA clones
dbVar	4,305,990	genome structural variation studies
Epigenomics	6,635	epigenomic studies and display tools
Genome	10,777	genome sequencing projects by organism
GSS	37,646,655	genome survey sequences
Nucleotide	155,458,773	DNA and RNA sequences
Probe	31,890,151	sequence-based probes and primers
SNP	444,458,769	short genetic variations
SRA	1,064,488	high-throughput DNA and RNA sequence read archive
Taxonomy	1,310,804	taxonomic classification and nomenclature catalog

Genes

EST	75,736,497	expressed sequence tag sequences
Gene	18,350,280	collected information about gene loci
GEO DataSets	1,332,018	functional genomics studies
GEO Profiles	108,708,851	gene expression and molecular abundance profiles
HomoloGene	141,268	homologous gene sets for selected organisms
PopSet	212,126	sequence sets from phylogenetic and population studies
UniGene	6,473,284	clusters of expressed transcripts

Proteins

Conserved Domains	49,955	conserved protein domains
Protein	153,454,195	protein sequences
Protein Clusters	820,546	sequence similarity-based protein clusters
Structure	103,644	experimentally-determined biomolecular structures

Chemicals

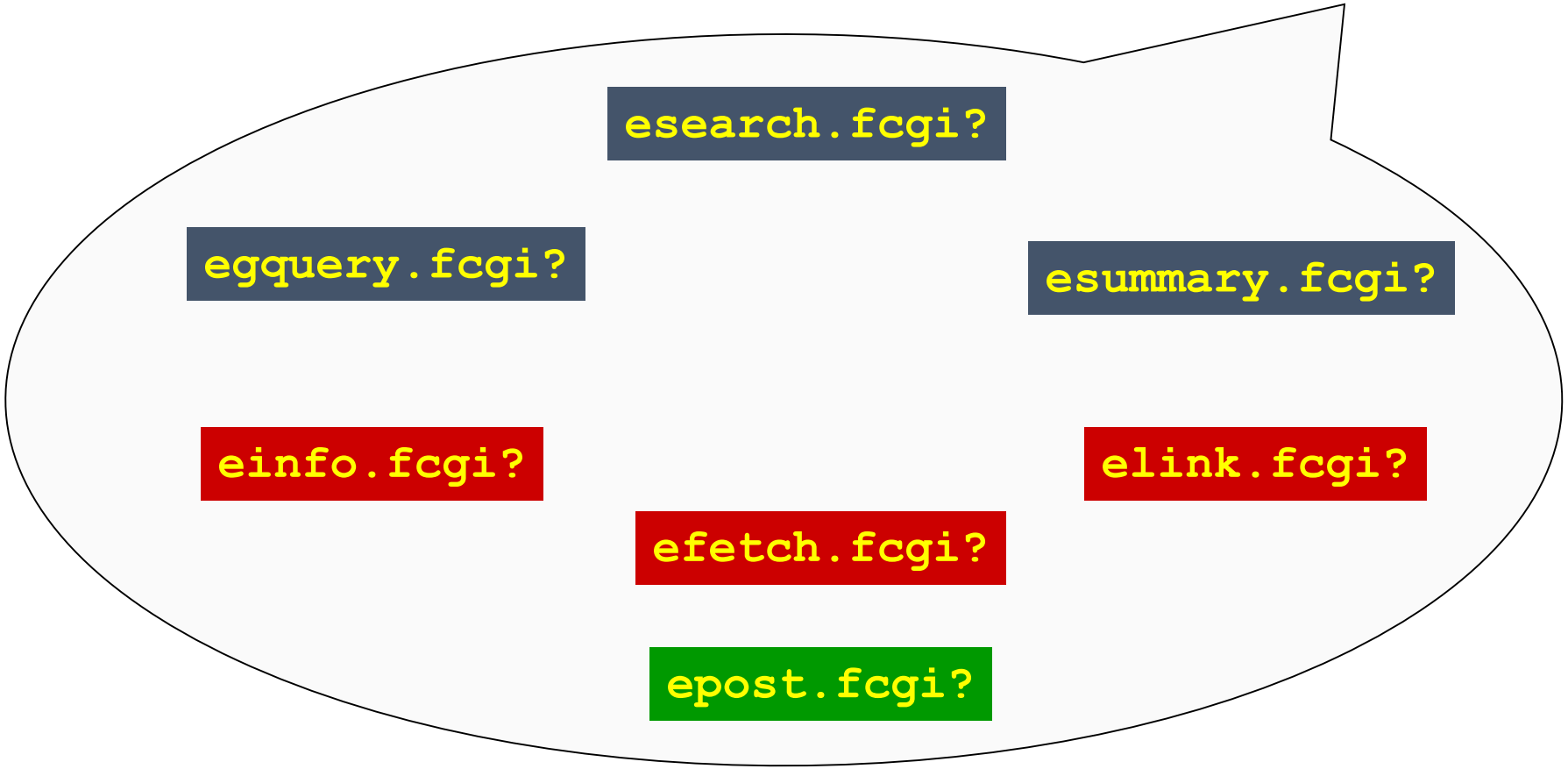
BioSystems	653,443	molecular pathways with links to genes, proteins and chemicals
PubChem BioAssay	1,112,105	bioactivity screening studies
PubChem Compound	62,041,347	chemical information with structures, information and links
PubChem Substance	177,330,151	deposited substance and chemical information

Entrez Programming Utilities (E-utilities)

The E-utilities are the public application programming interface (API) to the NCBI Entrez system and allow access to all Entrez databases including PubMed, PMC, Gene, Nuccore and Protein. The E-utilities are a suite of several **server-side programs** that accept a fixed URL syntax for search, link and retrieval operations.

The E-utilities base

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



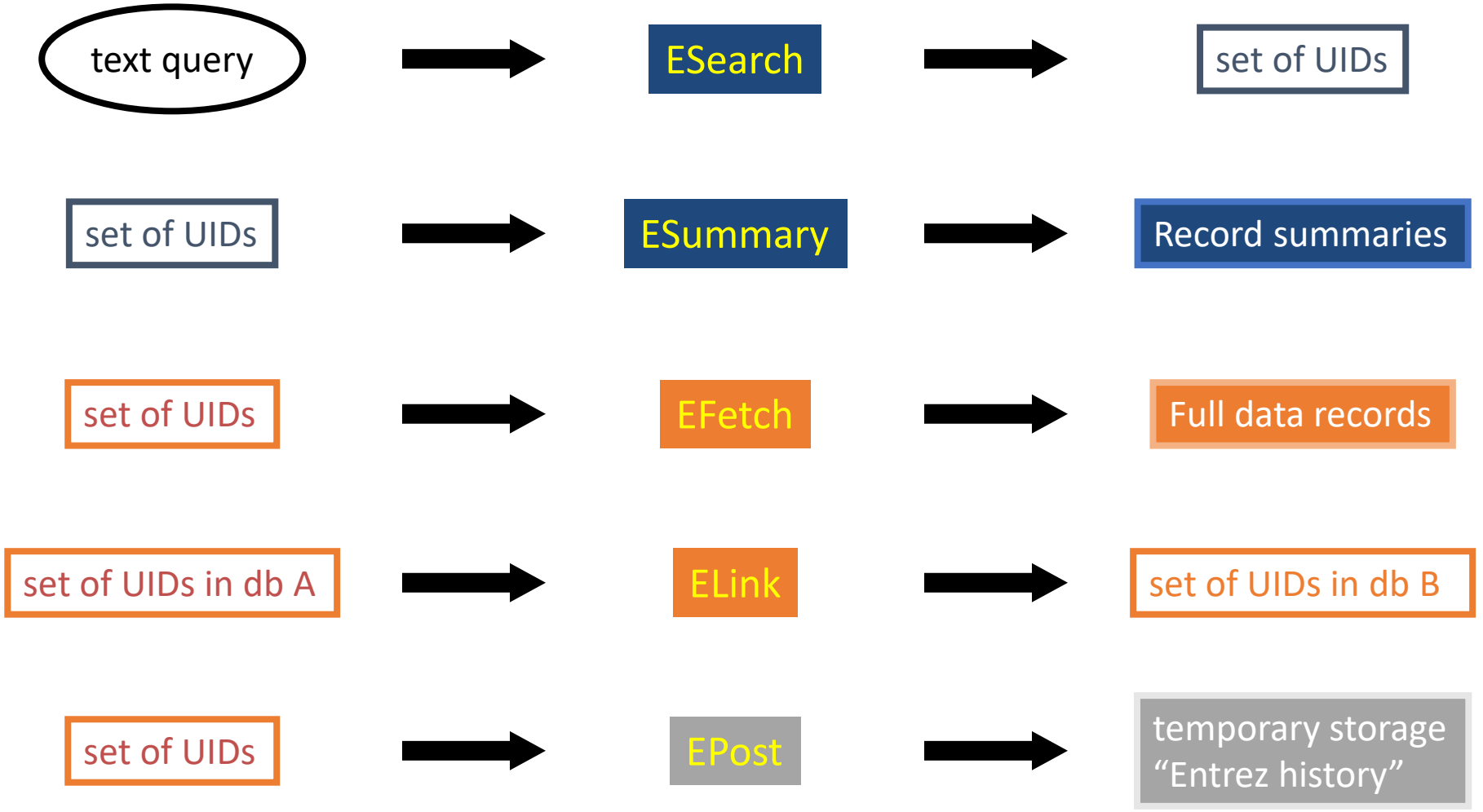
FastCGI is a binary protocol for interfacing interactive programs with a web server.

- http://eutils.ncbi.nlm.nih.gov/entrez/eutils/esearch.fcgi?db=pubmed&term=

Entrez database and Unique record IDentifiers (UIDs)

Entrez Database	UID common name	E-utility Database Name
BioProject	BioProject ID	bioproject
BioSample	BioSample ID	biosample
Biosystems	BSID	biosystems
Books	Book ID	books
Conserved Domains	PSSM-ID	cdd
dbGaP	dbGaP ID	gap
dbVar	dbVar ID	dbvar
Epigenomics	Epigenomics ID	epigenomics
EST	GI number	nucest
Gene	Gene ID	gene
Genome	Genome ID	genome
GEO Datasets	GDS ID	gds
GEO Profiles	GEO ID	geoprofiles
GSS	GI number	nucgss
HomoloGene	HomoloGene ID	homologene
MeSH	MeSH ID	mesh
NCBI C++ Toolkit	Toolkit ID	toolkit
NCBI Web Site	Web Site ID	ncbisearch
NLM Catalog	NLM Catalog ID	nlmcatalog
Nucleotide	GI number	nuccore
OMIA	OMIA ID	omia
PopSet	PopSet ID	popset
Probe	Probe ID	probe
Protein	GI number	protein
Protein Clusters	Protein Cluster ID	proteinclusters
PubChem BioAssay	AID	pcassay
PubChem Compound	CID	pccompound
PubChem Substance	SID	pcsubstance
PubMed	PMID	pubmed
PubMed Central	PMCID	pmc
SNP	rs number	snp
SRA	SRA ID	sra
Structure	MMDB-ID	structure
Taxonomy	TaxID	taxonomy
UniGene	UniGene Cluster ID	unigene
UniSTS	STS ID	unists

What do the E-utilities do?



Web Equivalents

Web action	E-utility equivalent
PubMed search (`maelstrom AND piRNA`)	**ESearch → ESummary**
Click on title in search results	EFetch
Click to full text	ELink
Gene search	**ESearch (→ ESummary/EFetch)**
Find all RefSeq proteins for the gene	ELink

The Basics: Esearch, Esummary and Efetch

PubMed Search

http://eutils.ncbi.nlm.nih.gov/entrez/eutils/esearch.fcgi?db=pubmed&term=maelstrom[All Fields] AND ("rna, small interfering"[MeSH Terms] OR ("rna"[All Fields] AND "small"[All Fields] AND "interfering"[All Fields]) OR "small interfering rna"[All Fields] OR "pirna"[All Fields])

PubMed ESummary

http://eutils.ncbi.nlm.nih.gov/entrez/eutils/esummary.fcgi?db=pubmed&id=25303775,25295037&version=2.0

PubMed EFetch

http://eutils.ncbi.nlm.nih.gov/entrez/eutils/efetch.fcgi?db=pubmed&id=25303775,25295037&rettype=abstract&retmode=text

Protein EFetch

http://eutils.ncbi.nlm.nih.gov/entrez/eutils/efetch.fcgi?db=nuccore&id=25303775,25295037&rettype=fasta&retmode=text

Entrez Programming Utilities Help

- https://www.ncbi.nlm.nih.gov/books/NBK25501/

Entrez Direct

- Command-line version of the NCBI Entrez API E-utilities
- A set of Perl scripts designed to be used as UNIX/Linux executables
- Each script outputs XML that can be piped directly into another script
- Requirements
  - UNIX, LINUX, Mac OSX
  - Perl with LWP::Simple
- Package contents
  - esearch
  - esummary
  - efetch
  - elink
  - epost
  - efilter (performs an esearch after an elink or esearch)
  - xtract (powerful XML parser)

Why use EDirect?

- Allows construction of custom pipelines for processing data
- Generates highly flexible custom output reports
- Built in batch access

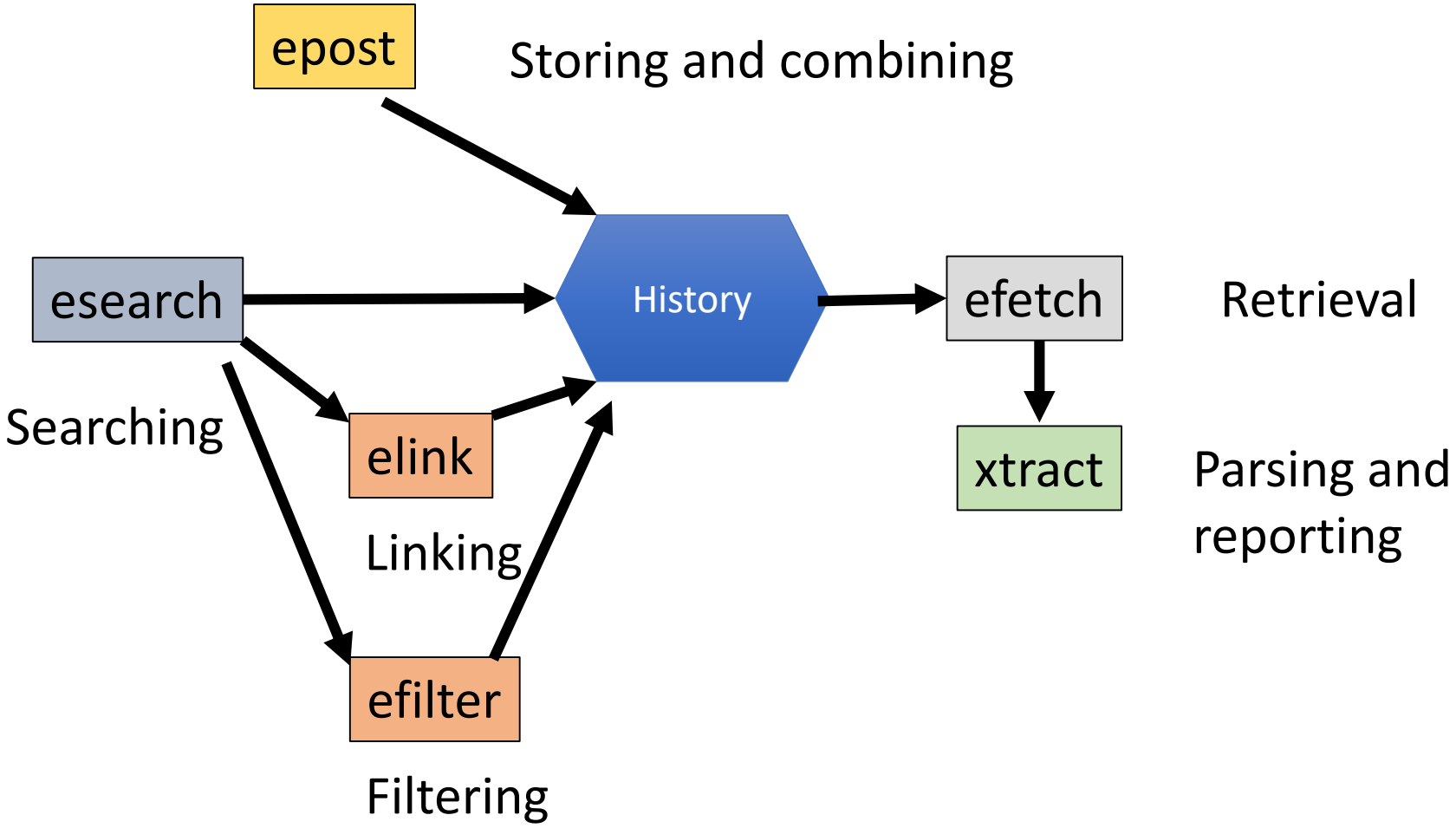
Goal: to learn the basics of using the EDirect programs to extract custom reports from the NLM/NCBI Literature and molecular databases

Installing EDirect

```
cd ~
  perl -MNet::FTP -e \
    '$ftp = new Net::FTP("ftp.ncbi.nlm.nih.gov", Passive => 1);
     $ftp->login; $ftp->binary;
     $ftp->get("/entrez/entrezdirect/edirect.zip");'
  unzip -u -q edirect.zip
  rm edirect.zip
  export PATH=$PATH:$HOME/edirect
  ./edirect/setup.sh
```

- Installs EDirect in your home directory and appends it to your PATH.
- Don’t do it for this class! EDirect is already installed in /usr/bin and is on everyone’s path on the instance.

EDirect workflows



einfo

Provides information about the available databases

- Available indexed fields
- Available links
- Produces XML (or text output with –fields, -links)

```
einfo -dbs
einfo -db dbname
einfo -db dbname -fields
einfo -db dbname -links
```

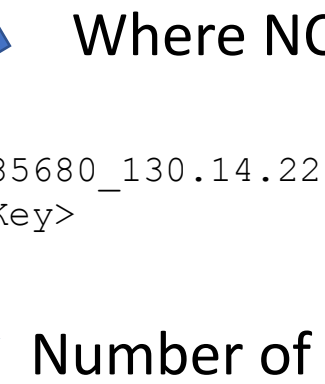
einfo –db gene

```
zhangd3@lmem21:~> einfo -db gene
<?xml version="1.0" encoding="UTF-8" ?>
<!DOCTYPE eInfoResult PUBLIC "-//NLM//DTD einfo 20130322//EN" "https://eutils.ncbi.nlm.nih.gov/eutils/dtd/201303
22/einfo.dtd">
<eInfoResult>
        <DbInfo>
        <DbName>gene</DbName>
        <MenuName>Gene</MenuName>
        <Description>Gene database</Description>
        <DbBuild>Build170821-0225m.1</DbBuild>
        <Count>29372491</Count>
        <LastUpdate>2017/08/21 09:38</LastUpdate>
        <FieldList>
                <Field>
                        <Name>ALL</Name>
                        <FullName>All Fields</FullName>
                        <Description>All terms from all searchable fields</Description>
                        <TermCount>459318739</TermCount>
                        <IsDate>N</IsDate>
                        <IsNumerical>N</IsNumerical>
                        <SingleToken>N</SingleToken>
                        <Hierarchy>N</Hierarchy>
                        <IsHidden>N</IsHidden>
                        <IsTruncatable>Y</IsTruncatable>
                        <IsRangable>N</IsRangable>
                </Field>
                <Field>
                        <Name>UID</Name>
                        <FullName>UID</FullName>
                        <Description>Unique number assigned to a gene record</Description>
                        <TermCount>0</TermCount>
                        <IsDate>N</IsDate>
                        <IsNumerical>Y</IsNumerical>
                        <SingleToken>Y</SingleToken>
                        <Hierarchy>N</Hierarchy>
                        <IsHidden>Y</IsHidden>
                        <IsTruncatable>N</IsTruncatable>
                        <IsRangable>Y</IsRangable>
```

esearch

- Uses standard web Entrez queries
  - try searches on web interface first
- Results stored in web environment
  - Pipe output to efetch, elink

```
esearch -db pubmed -query "maelstrom AND piRNA"
```



Where NCBI will store the results (history server)

```
<ENTREZ_DIRECT>
  <Db>pubmed</Db>
  <WebEnv>NCID_1_59385680_130.14.22.215_9001_1503366190_1402250482_0MetA0_S_MegaStore_F_1</WebEnv>
  <QueryKey>1</QueryKey>
  <Count>22</Count>
  <Step>1</Step>
</ENTREZ_DIRECT>
```

Number of the results

elink

- Returns related records in the same (-related) or different (-target) database
- Use link name from einfo to get the most precise results (-name linkname)
- Pipe into efetch
- Use with –cmd neighbor to get a table of linked identifiers (elink XML)

```
esearch -db pubmed -query "maelstrom AND piRNA" | elink -related


<ENTREZ_DIRECT>
  <Db>pubmed</Db>
  <WebEnv>NCID_1_59529507_130.14.18.34_9001_1503366443_873253847_0MetA0_S_MegaStore_F_1</WebEnv>
  <QueryKey>3</QueryKey>
  <Count>1838</Count>
  <Step>2</Step>
</ENTREZ_DIRECT>
```

```
esearch -db pubmed -query "maelstrom AND piRNA" | elink -target protein


<ENTREZ_DIRECT>
  <Db>protein</Db>
  <WebEnv>NCID_1_59551392_130.14.18.34_9001_1503366545_1925390075_0MetA0_S_MegaStore_F_1</WebEnv>
  <QueryKey>3</QueryKey>
  <Count>26</Count>
  <Step>2</Step>
</ENTREZ_DIRECT>
```

```
esearch -db pubmed -query "maelstrom AND piRNA“ |\
elink –related |\
elink -target protein
```

```
esearch -db pubmed -query "maelstrom AND piRNA" | elink -related | elink -target protein
<ENTREZ_DIRECT>
  <Db>protein</Db>
  <WebEnv>NCID_1_59576797_130.14.18.34_9001_1503366669_1979365779_0MetA0_S_MegaStore_F_1</WebEnv>
  <QueryKey>5</QueryKey>
  <Count>17171</Count>
  <Step>3</Step>
</ENTREZ_DIRECT>
```

```
elink -target structure
```

efilter

- uses the History server to filter or restrict the results of a previous query

```
esearch -db pubmed -query "maelstrom AND piRNA“ |\
efilter –query “structure[TIAB]”


<ENTREZ_DIRECT>
  <Db>pubmed</Db>
  <WebEnv>NCID_1_59838025_130.14.22.215_9001_1503368311_544618281_0MetA0_S_MegaStore_F_1</WebEnv>
  <QueryKey>2</QueryKey>
  <Count>5</Count>
  <Step>2</Step>
</ENTREZ_DIRECT>
```

Query Specification
-query Query string

Document Order
-sort Result presentation order

Date Constraint
-days Number of days in the past
-datetype Date field abbreviation
-mindate Start of date range
-maxdate End of date range

Publication Filters
-pub abstract, clinical, english, free, historical, journal, last_week, last_month, last_year, preprint, review, structured

Sequence Filters
-feature gene, mrna, cds, mat_peptide, ...
-location mitochondrion, chloroplast, plasmid, plastid
-molecule genomic, mrna, trna, rrna, ncrna
-organism animals, archaea, bacteria, eukaryotes, fungi, human, insects, mammals, plants, prokaryotes, protists, rodents, viruses
-source genbank, insd, pdb, pir, refseq, swissprot, tpa

Gene Filters
-status alive
-type coding, pseudo

Miscellaneous Arguments
-label Alias for query step

```
esearch -db pubmed -query "maelstrom AND piRNA" | elink -target protein
| efilter -organism insect

<ENTREZ_DIRECT>
  <Db>protein</Db>
  <WebEnv>NCID_1_59868754_130.14.18.34_9001_1503368022_325189468_0MetA0_S_MegaStore_F_1</WebEnv>
  <QueryKey>4</QueryKey>
  <Count>9</Count>
  <Step>3</Step>
</ENTREZ_DIRECT>
```

efetch/esummary

- Produces full XML records and Summaries (Docsums) for many databases
- Also specialized output for PubMed, sequence databases, Gene and others
- In many cases Docsums contain enough information (efetch –format docsum == esummary)
  - Parsing values from full XML can be more challenging

- efetch –help for supported return format

Efetch fully supported databases

- Literature databases
  - pubmed: biomedical abstracts
  - pmc: full text journal articles
  - mesh: MeSH ontology
  - nlmcatalog: NLM holdings
- Sequence databases
  - nuccore: DNA, RNA sequences
  - nucest: expressed sequence tags
  - protein: protein sequences
  - popset: population studies
  - sra: sequence read archive (next gen)
  - taxonomy: organisms
- Gene, variation, expression
  - gene: gene loci
  - homologene: homologous genes
  - snp: small variations
  - gds: expression databases
  - Biosamples: samples in other datasets
  - Biosystems: biological pathways

```
esearch -db pubmed -query "maelstrom AND piRNA“ |\
efetch –format abstract
```

1. Oncotarget. 2017 Jan 17;8(3):5026-5037. doi: 10.18632/oncotarget.13756.

Mael is essential for cancer cell survival and tumorigenesis through protection of genetic integrity.

Kim SH(1), Park ER(1), Cho E(1), Jung WH(1), Jeon JY(1), Joo HY(1), Lee KH(1), Shin HJ(1).

Author information:
(1)Division of Radiation Cancer Research, Korea Institute of Radiological & Medical Sciences, Seoul 139-706, Republic of Korea.

Germ line-specific genes are activated in somatic cells during tumorigenesis, and are accordingly referred to as cancer germline genes. Such genes that act on piRNA (Piwi-interacting RNA) processing play an important role in the progression of cancer cells. Here, we show that the spermatogenic transposon silencer maelstrom (Mael), a piRNA-processing factor, is required for malignant transformation and survival of cancer cells. A specific Mael isoform was distinctively overexpressed in diverse human cancer cell lines and its depletion resulted in cancer-specific cell death, characterized by apoptosis and senescence, accompanied by an increase in reactive oxygen-species and DNA damage. These biochemical changes and death phenotypes induced by Mael depletion were dependent on ATM. Interestingly Mael was essential for Myc/Ras-induced transformation, and its overexpression inhibited Ras-induced senescence. In addition, Mael repressed retrotransposon activity in cancer cells. These results suggest that Mael depletion induces ATM-dependent DNA damage, consequently leading to cell death specifically in cancer cells. Moreover, Mael possesses oncogenic potential that can protect against genetic instability.

DOI: 10.18632/oncotarget.13756
PMCID: PMC5354889
PMID: 27926513

```
esearch -db pubmed -query "maelstrom AND piRNA" | elink -target protein
| efetch -format fasta

>sp|Q9VF26.1|SPNE_DROME RecName: Full=Probable ATP-dependent RNA helicase spindle-E; AltName: Full=Homeless
MDQEVMDFFDFSKELKRVAAAPQGYISSDPRLMATKFKSSEVPNRELIGTDYVSKIVAKEKCLLNGTLLN
EQPQGKRIRTLDDLDTDDEGEETEIRRDDEYYKKFRFNLNRDKNLSIYAKREEILAAINAHPVVIIKGET
GCGKTTQVPQYILDEAYKSGKYCNIVVTQPRRIAAISIANRVCQEREWQQNTVCSFQVGLHRPNSLEDTR
LLYCTTGVLLNNLINNKTLTHYTHIVLDEVHERDQNMDFLLIVVRRLLATNSRHVKIILMSATIDAKELS
DYFTTTNSIPPVITTNHRRKHSIEKFYRDQLGSIIWNEEDVGHQQVPEINKHGYRAAVKIIVIIDNMERK
AAIQSRQSYDEALRYGAVLIFLPGIYEIDTMAENLTCMLENDPNIKVSIVRCFSLMTPENQRDVFNPPPP
GFRKIILTTNIAESSITVPDVSYVIDFCLAKVKVTDTASSFSSLRLTWASKANCRQRAGRVGRLRSGRVY
RMVNKHFYQREMPEFGIPEMLRLPLQNSVLKAKVLNMGSPVEILALALSPPNLSDIHNTILLLKEVGALY
LTVDGIYDPLDGDLTYWGTIMSRLPLDTRQSRLIILGYIFNMLEEAIIIAAGLSTPGLFAHEGGRSQLGD
SFWMHYIFSDGSGSDLVAIWRVYLTYLNIVENGHDQESAIRWAKRFHVSLRSLKEIHLLVQELRVRCTHL
GLIPFPVNPNQMMDDREKAIMLKVIIAGAFYPNYFTRSKESCADTDRNIYQTISGHDPCRTVYFTNFKPA
YMGELYTRRIKELFQEVRIPPENMDVTFQEGSQKVFVTFKQDDWIEGSSKYVPVSGRVQSEVYKAVMMRQ
NRVERPIHIMNPSAFMSYVQQRGIGDVIEGRWIPPTKPLNVELLALPSVFDKTISGSITCIVNCGKFFFQ
PQSFEECIRNMSEIFNAPQQLRNYVTNASAIAKGMMVLAKRDSYFQRATVIRPENQSNRQPMFYVRFIDY
GNCTLLPMQLMRLMPRELTEQYGDLPPRVFECRLAMVQPSSVVSGNNRWSTAANDMLKTVAQCGLIDIEV
YSLFNNVAAVLIHMRDGIINDKLVELMLCRRSDEDYMSRKDHDFRLRRQESARNLSTAQRQQINEEYLRS
CQLPQDHDLPPPPLEKCKTVVMLKGPNSPLECTMRSITRVGLSKRVNIDHLSVNALLLDADPQDHHDHLI
VAHEIAESRNGQTLTARGTTLMPNVQGFGALMVMLFSPTMQLKCNKEGTSYVSVLGGLGCDPDTNEPYFA
EHDVLINLDVNILEDDVILINQIRYYIDSVFFNFKEENNPAVSVNERVSIYTQLRSLINRLLCKDRRYIE
RNMSNADFEWETNPELPLPNEPFGKRAIFPMHSLTELQEEDTGRLVQLRENCSMLHKWRNFEGTLPHMTC
KLCNQLLESVPQLRLHLLTILHRDREKQIDYCNQ
>sp|Q14BI7.3|TDRD9_MOUSE RecName: Full=Putative ATP-dependent RNA helicase TDRD9; AltName: Full=Tudor domain-
containing protein 9
```

Sequence Database Formats

-format	-mode	Report Type
acc		Accession Number
est		EST Report
fasta		FASTA
fasta	xml	TinySeq XML
fasta_cds_aa		FASTA of CDS Products (Proteins)
fasta_cds_na		FASTA of Coding Regions
ft		Feature Table
gb		GenBank Flatfile
gb	xml	GBSet XML
gbc	xml	INSDSet XML
gbwithparts		GenBank with Contig Sequences
gene_fasta		FASTA of Gene

-format	-mode	Report Type
gp		GenPept Flatfile
gp	xml	GBSet XML
gpc	xml	INSDSet XML
gss		GSS Report
ipg		Identical Protein Report
ipg	xml	IPGReportSet XML
native	text	Seq-entry ASN.1
native	xml	Bioseq-set XML
seqid		Seq-id ASN.1

Structured Data--- XML format

- A makeup language:
  - HTML (hypertext markup language)
  - XML (extensible markup language): a tree-based or hierarchical structure of the data

- Advantage of XML is that many pieces of information are in specific locations in a well-defined data hierarchy

- Accessing individual units of data (value) that are fielded by name (descriptor), such as

```
<PubData>2013</PubData>
<Source>PLoS</Source>
<Volume>8</Volume>
<Issue>3</Issue>
<Pages>e58144</Pages>
```

xtract

- General Full-featured XML parser
- Produces tab delimited output
- Loops over XML structure using exploration options
- Prints out selected items (elements) from XML
- Conditional execution
- Flexible output formats

```
esearch -db pubmed -query "maelstrom AND piRNA" | efilter -query
"structure [TIAB]" | efetch -format docsum
```

```
<?xml version="1.0" encoding="UTF-8" ?>
<!DOCTYPE DocumentSummarySet PUBLIC "-//NLM//DTD esummary pubmed 20160808//EN"
"https://eutils.ncbi.nlm.nih.gov/eutils/dtd/20160808/esummary_pubmed.dtd">

<DocumentSummarySet status="OK">
<DbBuild>Build170819-2207m.6</DbBuild>

<DocumentSummary><Id>25865890</Id>
        <PubDate>2015 Apr 21</PubDate>
        <EPubDate>2015 Apr 9</EPubDate>
        <Source>Cell Rep</Source>
        <Authors>
                <Author>
                        <Name>Matsumoto N</Name>
                        <AuthType>Author</AuthType>
                        <ClusterID></ClusterID>
                </Author>
                <Author>
                        <Name>Sato K</Name>
                        <AuthType>Author</AuthType>
                        <ClusterID></ClusterID>
                </Author>
                <Author>
                        <Name>Nishimasu H</Name>
                        <AuthType>Author</AuthType>
                        <ClusterID></ClusterID>
                </Author>
                <Author>
                        <Name>Namba Y</Name>
                        <AuthType>Author</AuthType>
                        <ClusterID></ClusterID>
                </Author>
```

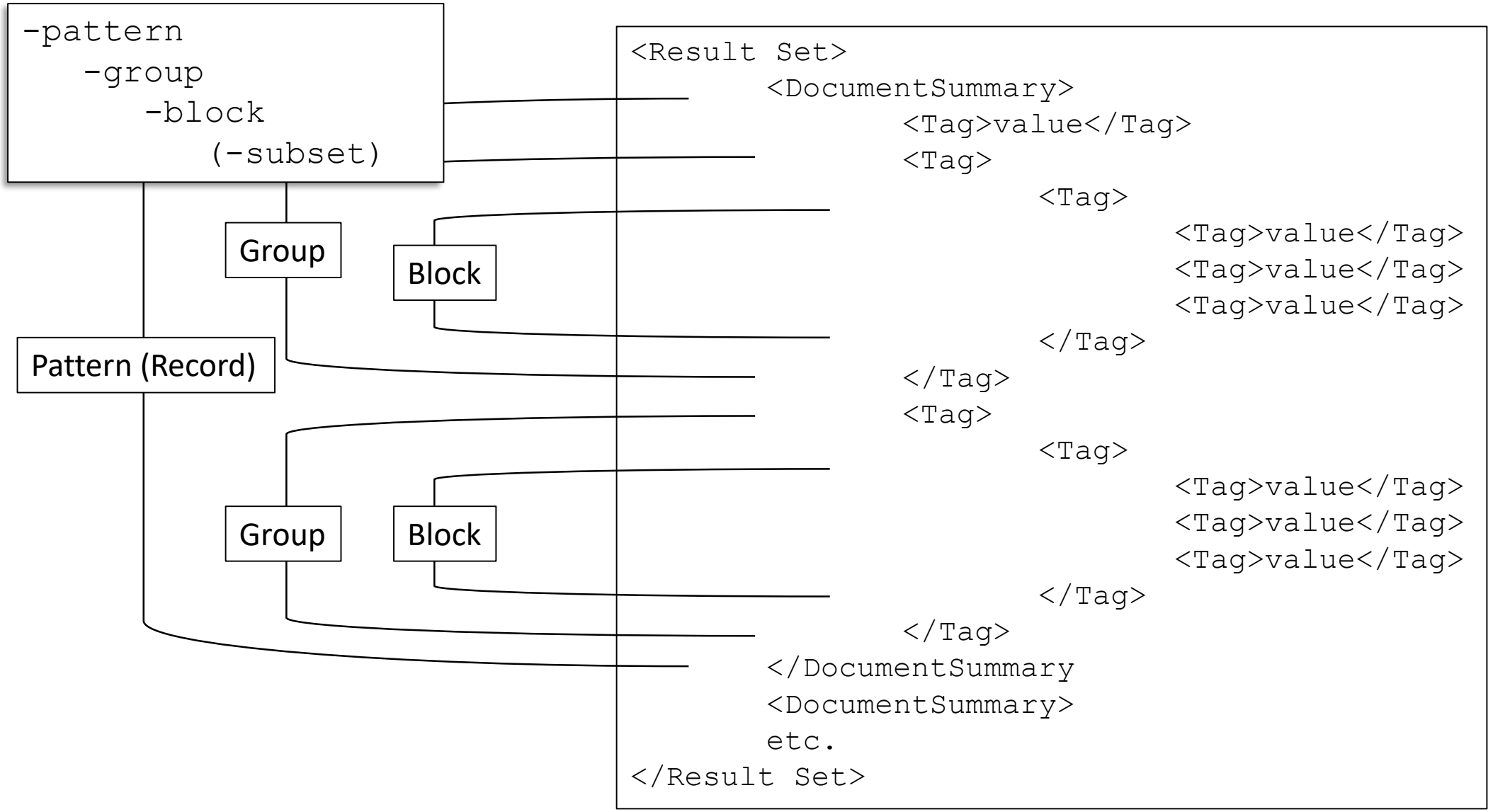
```
esearch -db pubmed -query "maelstrom AND piRNA" | efilter -query
"structure [TIAB]" | efetch -format docsum | xtract -outline

DbBuild
DocumentSummary
  Id
  PubDate
  EPubDate
  Source
  Authors
    Author
      Name
      AuthType
      ClusterID
    Author
      Name
      AuthType
      ClusterID
    Author
      Name
      AuthType
      ClusterID
  LastAuthor
  Title
  SortTitle
  Volume
  Issue
  Pages
  Lang
    string
  NlmUniqueID
  ISSN
```

The outline view presents a clear, uncluttered picture of the XML hierarchy that is useful in designing the appropriate command for data extraction.

How xtract parses XML

- -pattern places the data from individual records into separate rows.
- -element extracts values from specified fields into separate columns.
- -group, -block, and -subset limit element exploration to selected XML subregions.

Exploration options



```
esearch -db pubmed -query "maelstrom AND piRNA" | efilter -query
"structure [TIAB]" | efetch -format docsum | xtract  -pattern
DocumentSummary -element Id SortFirstAuthor Title


25865890        Matsumoto N     Crystal Structure and Activity of the Endoribonuclease Domain of the piRNA Pathway
25778731        Chen KM Metazoan Maelstrom is an RNA-binding protein that has evolved from an ancient nuclease active
23136393        Pek JW  Polo-mediated phosphorylation of Maelstrom regulates oocyte determination during oogenesis in
20362446        Patil VS        Repression of retroelements in Drosophila germline via piRNA pathway by the Tudor
17428915        Lim AK  Unique germ-line organelle, nuage, functions to repress selfish genetic elements in Drosophila
```

Sequence records

```
esearch -db pubmed -query "conotoxin" | elink -target protein |efilter -query
"mat_peptide[FKEY]"| efetch -format gpc | xtract -outline
```

```
INSDSeq
  INSDSeq_locus
  INSDSeq_length
  INSDSeq_moltype
  INSDSeq_topology
  INSDSeq_division
  INSDSeq_update-date
  INSDSeq_create-date
  INSDSeq_definition
  INSDSeq_primary-accession
  INSDSeq_accession-version
  INSDSeq_other-seqids
    INSDSeqid
    INSDSeqid
  INSDSeq_secondary-accessions
    INSDSecondary-accn
```

Those are the major elements.

efetch -db protein -id NP_001103824 -format gpc| xtract -insd complete mat_peptide "%peptide" product peptide

```
<INSDFeature>
    <INSDFeature_key>mat_peptide</INSDFeature_key>
    <INSDFeature_location>23..424</INSDFeature_location>
    <INSDFeature_intervals>
     <INSDInterval>
      <INSDInterval_from>23</INSDInterval_from>
      <INSDInterval_to>424</INSDInterval_to>
      <INSDInterval_accession>NP_001103824.1</INSDInterval_accession>
     </INSDInterval>
    </INSDFeature_intervals>
    <INSDFeature_quals>
     <INSDQualifier>
      <INSDQualifier_name>product</INSDQualifier_name>
      <INSDQualifier_value>zona pellucida sperm-binding protein 3 isoform 1</INSDQualifier_value>
     </INSDQualifier>
     <INSDQualifier>
      <INSDQualifier_name>calculated_mol_wt</INSDQualifier_name>
      <INSDQualifier_value>44386</INSDQualifier_value>
     </INSDQualifier>
     <INSDQualifier>
      <INSDQualifier_name>peptide</INSDQualifier_name>
```

Qualifier name

Qualifier value

<INSDQualifier_value>QPLWLLQGGASHPETSVQPVLVECQEATLMVMVSKDLFGTGKLIRAADLTLGPEACEPLVSMDTEDVVRFEVGLHECGNSMQVTDDALVYSTFLLHDPRPVGNLSIVR
TNRAEIPIECRYPRQGNVSSQAILPTWLPFRTTVFSEEKLTFSLRLMEENWNAEKRSPTFHLGDAAHLQAEIHTGSHVPLRLFVDHCVATPTPDQNASPYHTIVDFHGCLVDGLTDASSAFKVPRPG
PDTLQFTVDVFHFANDSRNMIYITCHLKVTLAEQDPDELNKACSFSKPSNSWFPVEGSADICQCCNKGDCGTPSHSRRQPHVMSQWSRSASRNRRHVTEEADVTVGPLIFLDRRGDHEVEQWA
LPSDTSVVLLGVGLAVVVSLTLTAVILVLTRRCRTASHPVSASE</INSDQualifier_value>

```
esearch -db pubmed -query "conotoxin" | elink -target protein |efilter -query "mat_peptide[FKEY]"| efetch -format gpc | xtract -insd complete mat_peptide "%peptide" product peptide
```

```
esearch -db pubmed -query "conotoxin" | elink -target protein |efilter -query "mat_peptide[FKEY]"| efetch -format gpc | xtract -insd source organism strain
```

```
NP_001103824.1  Homo sapiens    -
NP_036964.3     Rattus norvegicus       Sprague-Dawley
NP_683746.1     Mus musculus    C57BL/6
NP_775304.1     Mus musculus    C57BL/6
NP_000713.2     Homo sapiens    -
NP_065135.2     Homo sapiens    -
NP_113955.1     Rattus norvegicus       BN
NP_067344.2     Mus musculus    -
NP_072108.1     Rattus norvegicus       Sprague-Dawley
NP_062170.1     Rattus norvegicus       -
NP_075219.1     Rattus norvegicus       Sprague-Dawley
NP_072161.1     Rattus norvegicus       Sprague-Dawley
NP_434692.2     Rattus norvegicus       Sprague-Dawley
```

Get protein sequences from nucleotide accessions

```
cat accs_file | epost -db nuccore -format acc | elink -target protein | efetch -format fasta
```

Finding genes on chromosome Y

esearch -db gene -query "Homo Sapiens [ORGN] AND Y[CHR]” | efilter -query "alive [PROP]" | esummary | xtract -pattern DocumentSummary -NAME Name -block GenomicInfoType -match "ChrLoc:Y" -tab "\n" -element "&NAME", ChrAccVer,ChrStart,ChrStop

SRY	NC_000024.10	2787740	2786854
SHOX	NC_000024.10	624343	659410
CD99	NC_000024.10	2691132	2741308
IL3RA	NC_000024.10	1336574	1382688
CSF2RA	NC_000024.10	1268799	1325096
TSPY1	NC_000024.10	9466954	9469755
DAZ1	NC_000024.10	23199116	23129354
CRLF2	NC_000024.10	1212761	1190436
VAMP7	NC_000024.10	57067799	57130288
IL9R	NC_000024.10	57184100	57199536
SLC25A6	NC_000024.10	1392145	1386151
ASMT	NC_000024.10	1595454	1643080
DDX3Y	NC_000024.10	12903998	12920477
RBMY1A1	NC_000024.10	21534878	21559682
PCDH11Y	NC_000024.10	5000043	5742227
KDM5D	NC_000024.10	19745346	19692494
DAZ2	NC_000024.10	23219456	23291355
USP9Y	NC_000024.10	12701230	12860843
ZFY	NC_000024.10	2934401	2982507

- EDirect Example 1: ESearch (gene) - ELink (gene to protein) - EFetch (protein FASTA)
- --
- esearch -db gene -query "foxp2[gene] AND human[orgn] AND alive[prop]" | \
- elink -target protein -name gene_protein_refseq | \
- efetch -format fasta
- --

- EDirect Example 2: ESearch (gene) - ELink (gene to protein); each output line includes the gene ID in column 1 followed by the protein GI's linked to that gene
- --
- esearch -db gene -query "foxp2[gene] AND human[orgn] AND alive[prop]" | \
- elink -target protein -name gene_protein_refseq -cmd neighbor | \
- xtract -pattern LinkSet -block IdList -element Id -block LinkSetDb -element Id
- --

- EDirect Example 3: ESearch (gene) - ELink (gene to protein) - ESummary (protein); uses xtract to write a table including the accession (caption), sequence length (Slen) and title (Title)
- --
- esearch -db gene -query "foxp2[gene] AND human[orgn] AND alive[prop]" | \
- elink -target protein -name gene_protein_refseq | \
- esummary | \
- xtract -pattern DocumentSummary -element Caption Slen Title
- --

- EDirect Example 4: Same as example 3 except that the output is piped to UNIX sort to sort the output by decreasing sequence length
- --
- esearch -db gene -query "foxp2[gene] AND human[orgn] AND alive[prop]" | \
- elink -target protein -name gene_protein_refseq | \
- esummary | \
- xtract -pattern DocumentSummary -element Caption Slen Title | \
- sort -t $'\t' -k 2,2nr

Entrez Programming Utilities Help

- https://www.ncbi.nlm.nih.gov/books/NBK25501/