A Teratological Tour of NCBI

Eric W. Sayers, Ph.D
National Center for Biotechnology Information
June 30, 2005


Course Topics

• Overview of NCBI
• Resources for genetic disorders
  – OMIM: human disease phenotypes
  – From sequences to annotated genes
  – Microarrays and polymorphisms
  – Test case: Muenke syndrome
• Resources for chemical teratogens
  – PubChem datasets
  – Test case: 6-aminonicotinamide

The National Institutes of Health
The National Center for Biotechnology Information

- Created as a part of NLM in 1988
  - Establish public databases
  - Perform research in computational biology
  - Develop software tools for sequence analysis
  - Disseminate biomedical information

What does NCBI do?

- accepts submissions of primary data
- develops tools to analyze these data
- uses these tools to create derivative databases based on the primary data
- provides free search, link, and retrieval of these data, primarily through the Entrez system

Types of Databases

- Primary Databases
  - Original submissions by experimentalists
  - Content controlled by the submitter

- Derivative Databases
  - Built from primary data
  - Content controlled by third party (NCBI)
BLAST: Sequence Similarity Searches

VAST: Structure Similarity Searches

NCBI Web Traffic

Christmas and New Year's Day
Useful Databases for Teratology

- OMIM (Online Mendelian Inheritance in Man)
  - literature catalog of human disease phenotypes
- Gene
  - curated database of gene loci and associated sequences
- GEO (Gene Expression Omnibus)
  - primary and derivative collections of gene expression data
- SNP (Single Nucleotide Polymorphisms)
  - primary and derivative collections of nucleotide polymorphisms
- PubChem (Substance, Compound, Bioassay)
  - primary and derivative collections of small molecules and their biological activities

Test Cases: Muenke Syndrome and 6-AN

Muenke syndrome, or Muenke nonsyndromic coronal craniosynostosis, is a developmental disorder characterized by a premature closing of cranial sutures, resulting in deformities in the head and face. Muenke syndrome is caused by a mutation in the gene for fibroblast growth factor receptor 3, FGFR3.

Mutations in a related gene, fibroblast growth factor receptor 2, cause a similar craniosynostosis called Crouzon syndrome, along with several other related developmental problems.

6-AN, or 6-aminonicotinamide, is an antimetabolite of nicotinic acid (vitamin B-3) and is a known teratogen, causing lesions in the CNS, skin, and digestive tract during embryonic development.

6-AN has been linked to apoptosis in tumors, and is being investigated as a chemotherapeutic agent.
Searching in OMIM

NCBI Resources for Biomedical Research

muenke syndrome 33 records retrieved

muenke[All Fields] AND syndrome[All Fields]

"muenke syndrome" 2 records retrieved

"muenke syndrome"[All Fields]

muenke syndrome[title] 1 record retrieved

muenke syndrome[title]

OMIM

OMIM

Search Fields:

access to search fields

access to previous searches

query translation

An OMIM Record

An OMIM Record
OMIM Allelic Variant in FGFR3

OMIM Record for FGFR3

What is a Gene?
**Gene Record for FGFR3**

<table>
<thead>
<tr>
<th>Transcript and products: Reference below</th>
</tr>
</thead>
<tbody>
<tr>
<td>FGFR3 (FGF receptor 3)</td>
</tr>
</tbody>
</table>

Genomic context: chromosome 4: Map: 4q16.3

**Entrez Gene and RefSeq**

- Entrez Gene is the central depository for information about a gene available at NCBI, and often provides links to sites beyond NCBI
- Entrez Gene includes records for organisms that have NCBI Reference Sequences (RefSeqs)
- Entrez Gene records contain RefSeq mRNAs, proteins, and genomic DNA (if known) for a gene locus, plus links to other Entrez databases
- NCBI RefSeqs are based on primary sequence data in GenBank

**What is GenBank?**

NCBI’s Primary Sequence Database

- Nucleotide only sequence database
- Archival in nature
- Each record is assigned a stable accession number
- GenBank Data
  - Direct submissions (traditional records)
  - Batch submissions (EST, GSS, STS)
  - ftp accounts (genome data)
- Three collaborating databases
  - GenBank
  - DNA Database of Japan (DDBJ)
  - European Molecular Biology Laboratory (EMBL) Database
NCBI Resources for Biomedical Research

The International Sequence Database Collaboration

NIH

Entrez

Sequin
BankIt
ftp
Submissions
Updates

GenBank

EMBL

CIB

getentry

NIG

SRS

EBI

DDBJ

• Submissions
• Updates

Submissions
Submissions

The Growth of GenBank

Release 148: 45.2 million records
49.4 billion nucleotides

Average doubling time = 14 months*


Release 148 June 2005

45,236,251 Records
49,398,852,122 Nucleotides
>140,000 Species
172 Gigabytes 785 files

• full release every two months
• incremental and cumulative updates daily
• available only through internet
Primary vs. Derivative Databases

RefSeq: NCBI’s Derivative Sequence Database

RefSeq Benefits
- Non-redundant
- Explicitly linked nucleotide and protein sequences
- Updated to reflect current sequence data and biology
- Validated by hand
- Format consistency
- Distinct accession series
- Stewardship by NCBI staff and collaborators


RefSeq: NCBI’s Derivative Sequence Database

- Curated transcripts and proteins
  - NM_123456 → NP_123456
  - NR_123456 (non-coding RNA)
- Model transcripts and proteins
  - XM_123456 → XP_123456
  - XR_123456 (non-coding RNA)
- Assembled Genomic Regions (contigs)
  - NT_123456 (BAC clones)
  - NW_123456 (WGS)
- Other Genomic Sequence
  - NG_123456 (complex regions, pseudogenes)
  - NZ_ABCD12345678 (WGS) → ZP_123456
- Chromosome records in Entrez Genome
  - NC_123456 (chromosome, microbial or organelle genome)
NM/NP Records for FGFR3 in Entrez

NM_000142: variant 1

NM_022965: variant 2

mRNA that completes 3' end

RefSeq Summary for FGFR3

The RefSeqs for a gene become the basis for feature annotations such as SNPs

Searching in Gene

muenke syndrome 1 record retrieved

muenke[All Fields] AND syndrome[All Fields] AND alive[prop]

fgfr3 19 records retrieved

fgfr3[gene name] 8 records retrieved

fgfr3[gene name] AND human[orgn] 1 record retrieved

fgfr3[gene name] AND "Homo sapiens"[Organism] AND alive[prop]
NCBI Resources for Biomedical Research

GPL Platform descriptions
GSM Raw/processed spot intensities from a single slide/chip
GSE Grouping of slide/chip data "a single experiment"
GDS Grouping of experiments

Entrez GEO Profiles
Entrez GEO Datasets

Linking to GEO Profiles

OMIM

GEO Profiles for FGFR3

Each record represents one spot on one chip
NCBI Resources for Biomedical Research

Viewing a GEO Profile

- Red bars indicate raw spot intensities.
- Blue points indicate percentile rankings for that spot relative to all spots on the chip.

NCBI Resources for Biomedical Research

Viewing the GDS and GSE Records

- GEO Analysis Tools

14
Limit Results to FGFR3 with Entrez

NCBI Resources for Biomedical Research

Sequence Polymorphisms

- Clinical literature database
- Curated at Johns Hopkins Univ
- Links human genes and genetic disorders to human disease
- Lists allelic variants that have clinical consequences

Variations in SNP are not necessarily in OMIM, and vice versa!

Linking to SNP

The linking of publications are a limited number of publications which describe a known number of alleles, including microsatellites and insertions. They also include all other known translocations, insertions, and deletions. Translocations may involve translocations, insertions, and deletions.
SNP Links

**SNP**
List of all SNPs associated with the given record

**GeneView in dbSNP**
Graphical view of all SNPs associated with the given gene locus

**Gene Genotype**
List of populations and statistics of observed SNPs associated with the given record

---

SNP Links for FGFR3

---

SNP rs4647924

---

NCBI Resources for Biomedical Research
Genotype Detail

Click to view the pedigree chart

Pedigree Chart

Finding Genotype Data for a SNP


Nonsynonymous SNPs with genotype data in human FGFR2
### Population Data for rs755793

**Variation Summary:**
- Total number of genotypes with known allele frequency: 3
- Total number of genotypes with genotype data: 3
- Average allele frequency: 0.833

<table>
<thead>
<tr>
<th>Subjects</th>
<th>European</th>
<th>Asian</th>
</tr>
</thead>
<tbody>
<tr>
<td>Samples</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Healthy</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Click to access genotype detail and pedigree charts.

**LD Plots for FGFR2**

#### European

#### African

#### LD Plots for FGFR2

#### Asian

#### African
Genotype Detail for rs755793

Entrez PubChem

Search for a Teratogen: 6-AN
NCBI Resources for Biomedical Research

Results from PC Substance

A redundant set with multiple records per compound

CID (Compound ID) 9500

Compound Summary:

Redundant Substances

Bioassays

Search by structure

Similar compounds

Links to MeSH
Histogram of Activity Data

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Active Compounds in Bin #5

Transferring Data to Entrez
Finding Teratogens in the Active List

A More Active Teratogen

For More Information…

E-mail addresses

- General Help: info@ncbi.nlm.nih.gov
- BLAST: blast-help@ncbi.nlm.nih.gov

The (free!) NCBI Newsletter

http://www.ncbi.nih.gov/About/newsletter.html

The NCBI Handbook

Follow the link from the NCBI Home Page

The NCBI Education Page