1

Bioinformatic Methods I

Welcome to Bioinformatic Methods I*!*

Instructor: **Nicholas Provart**

Nicholas Provart is a professor in the Department of Cell & Systems Biology at the University of Toronto. He's taught a course on which this Coursera course is based since 2009 to approximately 1200 undergraduate University of Toronto biology students. His involvement with bioinformatics goes back to 1998. He was Director of the Collaborative Graduate Program in Genome Biology & Bioinformatics from 2006-2011, and is one of the founding members of the International Arabidopsis Informatics Consortium.

Please use the Coursera tools to discuss lecture content and labs.

Course material developed by Ryan Austin, David Guttman, Laura Hug, Momoko Price, and Nicholas Provart
Course produced by Jamie Waese, Rohan Patel, William Heikoop, and Nicholas Provart

2

Course format and syllabus

This Coursera course will cover the basics of searching one of the main repositories of sequence information, NCBI's GenBank, using GQuery/Entrez and Blast (Basic Local Alignment Search Tool), along with creating sequence alignments and phylogenies. Selection analysis will also be covered, as will next generation sequence analysis and metagenomics. Most tools used for exploration are web-based.

Week	Topic
1	**NCBI/Blast I**
2	Blast II/Comparative Genomics
3	Multiple Sequence Alignments
4	Phylogenetics
5	Selection Analysis
6	NGS Analysis / Metagenomics

Bioinformatic Methods II

Week	Topic
1	Protein motifs
2	Protein-protein interactions
3	Protein Structure
4	Gene Expression Analysis I
5	Gene Expression Analysis II
6	Cis regulatory elements

The weekly material will consist of video mini-lectures (20 minutes) and short (2 minute) intro and summary videos, weekly labs (1-2 hours) with lab quizzes (plus optional lab discussion videos), two section quizzes (one after the first 3 weeks, and the other at the end of the course), and one assignment (due at the end of the course).

What is bioinformatics?

Bioinformatics

- is the development and application of computational tools in managing all kinds of biological data
- involves the technology that uses computers for storage, retrieval, manipulation, and distribution of information related to biological macromoleculates such as DNA, RNA, proteins and metabolites
- generally limited to sequence, structural, and functional analysis of genes and genomes and their corresponding products
- sometimes called computational molecular biology

This field has developed over the past decade or so to help manage the huge increase in data generated by genome sequencing projects, high-throughput technologies etc.

Xiong (2006) Essential Bioinformatics, Cambridge University Press.

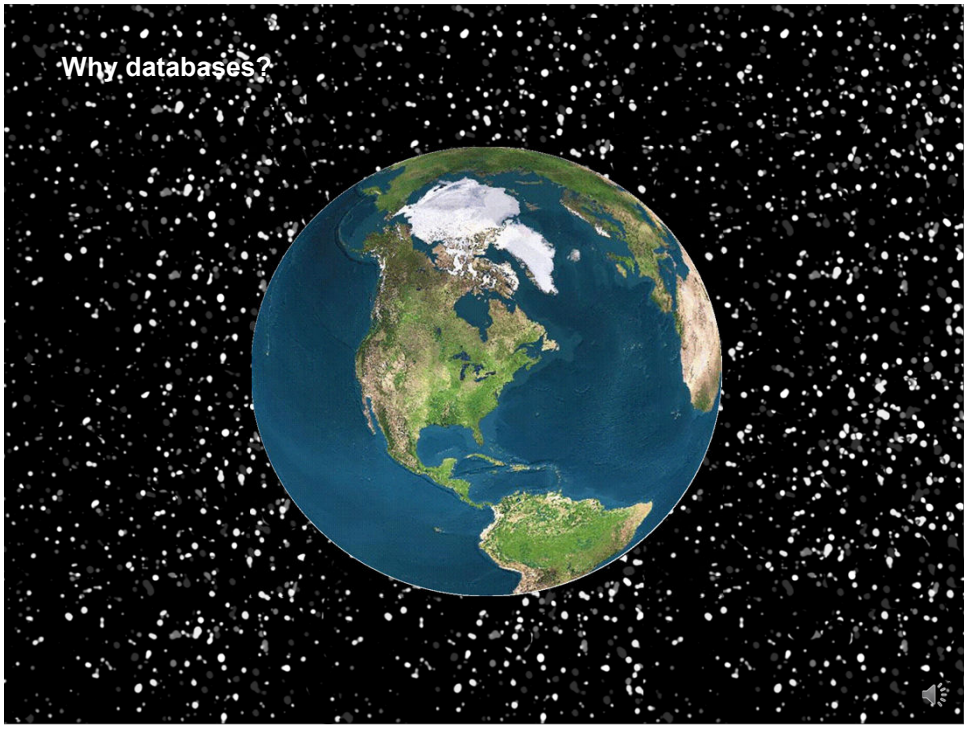
Why bioinformatics?

```
>gi|27500381:c623297-542205 Homo sapiens chromosome 17 genomic contig
AAAACTGCGACTGCGCGGCGTGAGCTCGCTGAGACTTCCTGGACGGGGGACAGGCTGTGGGGTTTCTCAG
ATAACTGGGCCCCTGCGCTCAGGAGGCCTTCACCCTCTGCTCTGGGTAAAGGTAGTAGAGTCCCGGGAAA
GGGACAGGGGGCCCAAGTGATGCTCTGGGGTACTGGCGTGGGAGAGTGGATTTCCGAAGCTGACAGATGG
GTATTCTTTGACGGGGGGTAGGGGCGGAACCTGAGAGGCGTAAGGCGTTGTGAACCCTGGGGAGGGGGGC
AGTTTGTAGGTCGCGAGGGAAGCGCTGAGGATCAGGAAGGGGGCACTGAGTGTCCGTGGGGGAATCCTCG
TGATAGGAACTGGAATATGCCTTGAGGGGGACACTATGTCTTTAAAAACGTCGGCTGGTCATGAGGTCAG
GAGTTCCAGACCAGCCTGACCAACGTGGTGAAACTCCGTCTCTACTAAAAATACAAAAATTAGCCGGGCG
TGGTGCCGCTCCAGCTACTCAGGAGGCTGAGGCAGGAGAATCGCTAGAACCCGGGAGGCGGAGGTTGCAG
TGAGCCGAGATCGCGCCATTGCACTCCAGCCTGGGCGACAGAGCGAGACTGTCTCAAAACAAAACAAAAC
AAAACAAAACAAAAAACACCGGCTGGTATGTATGAGAGGATGGGACCTTGTGGAAGAAGAGGTGCCAGGA
ATATGTCTGGGAAGGGGAGGAGACAGGATTTTGTGGGAGGGAGAACTTAAGAACTGGATCCATTTGCGCC
ATTGAGAAAGCGCAAGAGGGAAGTAGAGGAGCGTCAGTAGTAACAGATGCTGCCGGCAGGGATGTGCTTG
AGGAGGATCCAGAGATGAGAGCAGGTCACTGGGAAAGGTTAGGGGCGGGGAGGCCTTGATTGGTGTTGGT
TTGGTCGTTGTTGATTTTGGTTTTATGCAAGAAAAAGAAAACAACCAGAAACATTGGAGAAAGCTAAGGC
TACCACCACCTACCCGGTCAGTCACTCCTCTGTAGCTTTCTCTTTCTTGGAGAAAGGAAAAGACCCAAGG
GGTTGGCAGCAATATGTGAAAAAATTCAGAATTTATGTTGTCTAATTACAAAAAGCAACTTCTAGAATCT
TTAAAAATAAAGGACGTTGTCATTAGTTCTTTGGTTTGTATTATTCTAAAACCTTCCAAATCTTAAATTT
ACTTTATTTTAAAATGATAAAATGAAGTTGTCATTTTATAAACCTTTTAAAAAGATATATATATATGTTT
TTCTAATGTGTTAAAGTTCATTGGAACAGAAAGAAATGGATTTATCTGCTCTTCGCGTTGAAGAAGTACA
AAATGTCATTAATGCTATGCAGAAAATCTTAGAGTGTCCCATCTGGTAAGTCAGCACAAGAGTGTATTAA
TTTGGGATTCCTATGATTATCTCCTATGCAAATGAACAGAATTGACCTTACATACTAGGGAAGAAAAGAC
ATGTCTAGTAAGATTAGGCTATTGTAATTGCTGATTTCCTTAACTGAAGAACTTTAAAAATATAGAAAAT
GATTCCTTGTTCTCCATCCACTCTGCCTCTCCCACTCCTCTCCTTTTCAACACAAATCCTGTGGTCCGGG
AAAGACAGGGACTCTGTCTTGATTGGTTCTGCACTGGGGCAGGAATCTAGTTTAGATTAACTGGCATTTT
GGCTTTTCTTCCAGCTCTAAAACAAGCTCCATCACTTGAAATGGCAAAATAAAATCATGGATGAGGCCGA
GGGCGGTGGCTTATGCCTGTAATCCCAGCACTTTGGGAGGCCAAGGTGGTAGGATCACGAGGTCAGGAGA
TCGAGACCATCCTGGCCAACATGGTGAAACCCCCTCTCCACTAAAAATACAAAAATTAGCTGGGCGTAGT
```


Biological Databases

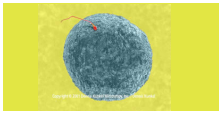
Outline

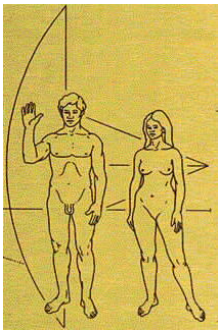
- Why databases?
- What is a database?
- Data structures: Flat File and Relational
- Accession numbers and identifiers
- A practical example of utility – NCBI Search (GQuery/Entrez)

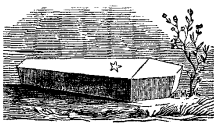


7

Why databases?







Genome and genomic sequences
Gene sequences, mutations
Gene regulation
Gene expression (where and when)
Intron splice variants
Protein sequence, post-translational modifications
Protein tertiary structure (3D)
Protein networks
Protein localization
Enzyme Kinetics
Metabolites, metabolic networks
Diseases
Literature

→To archive accumulated knowledge and to provide scientists with easy access to biological data

8

What is a database?

How can data be stored...

Flat-file format, with fields separated by some delimiter

Nancy|Dengler|Botany|University of Toronto|25 Willocks St, Toronto, ON. M5S 3B2
Peter|Lewis|Dept. of Biochemistry|Uni. Toronto|1 King's College Circle, Toronto, ON. M5S 1A8
John|Coleman|Department of Botany|University of Toronto|25 Willcocks St, Toronto, ON. M5S 3B2
John|Coleman|Dept. of Biology|York University|4700 Keele St, Toronto, ON. M3J 1P3

These data could also be stored in a spreadsheet

First_name	Last_name	Institution	Department	Address
Nancy	Dengler	University of Toronto	Botany	25 Willocks St, Toronto, ON. M5S 3B2
Peter	Lewis	Uni. Toronto	Dept. of Biochemistry	1 King's College Circle, Toronto, ON. M5S 1A8
John	Coleman	University of Toronto	Department of Botany	25 Willcocks St, Toronto, ON. M5S 3B2
John	Coleman	York University	Dept. of Biology	4700 Keele St, Toronto, ON. M3J 1P3

What are the problems with this sort of database?

Relational databases offer a solution...

Relational Databases

Nancy|Dengler|Botany|University of Toronto|25 Willocks St, Toronto, ON. M5S 3B2
Peter|Lewis|Dept. of Biochemistry|Uni. Toronto|1 King's College Circle, Toronto, ON. M5S 1A8
John|Coleman|Department of Botany|University of Toronto|25 Willcocks St, Toronto, ON. M5S 3B2
John|Coleman|Dept. of Biology|York University|4700 Keele St, Toronto, ON. M3J 1P3

A relational database consists of a relations (tables) containing attributes (fields or columns). Each row in a table is known as a tuple or a record. Information should be 'normalized' so that it is non-redundant → this means that every row should be unique, although this ideal is not always observed.

Table 'Professors'

Professor_id	First_name	Last_name	Contact_id
1	Nancy	Dengler	1
2	Peter	Lewis	2
3	John	Coleman	1
4	John	Coleman	3

primary key

foreign key

Table 'Contacts'

Contact_id	Institution	Department	Address
1	University of Toronto	Dept. of Botany	25 Willocks St, Toronto, ON. M5S 3B2
2	Uni. Toronto	Dept. of Biochemisty	1 King's College Circle, Toronto, ON. M5S 1A8
3	York University	Dept. of Biology	4700 Keele St, Toronto, ON. M3J 1P3

primary key

Accession codes, identifiers etc.

Many of the biolological databases (GenBank, UNIPROT etc.) have two (or more!) different ways of identifying a given entry:

- Identifier
- Accession code (or number)

Accession codes, identifiers etc. [2]

Identifier

An **identifier** ("locus" in GenBank, "entry name" in UNIPROT) is a string of letters and digits that might be understandable in some meaningful way by a human.

Identifiers are not as stable as accession numbers, mainly because they are modified by the curators if the presumed function of the protein is found to be something else.

UNIPROT: **`ADH6_HUMAN`**
GenBank: **`AH001409`** (formerly **`HUMADH6A01`**)

An identifier can change. For example, the database curators may decide that the identifier for an entry no longer is appropriate. This does not happen very often, but has actually happened recently for our example (used to be **`HUMADH6A01`** and more recently **`SEG_HUMADH6A0`** – now it seems GenBank has decided to use the same code for both locus/accession…)

Accession codes, identifiers etc. [3]

Accession code (number)

An **accession code** (or number) is a number (with a few characters in front) that uniquely identifies an entry. It is often assigned arbitrarily. For example, the accession code for ADH6_HUMAN in UNIPROT is P28332.

In the case of GenBank, the accession code for the human ADH6 gene sequence is AH001409.

Accession codes, identifiers etc. [4]

Versioning of sequences in GenBank

Records typically contain the **Accession.Version** identifier, such as AH001409.2, in the VERSION line of the record. This identifier used to be mapped to its corresponding GI* number, which was like the “primary key” of GenBank.

To specify **a sequence exactly** in GenBank, use its *Accession.Version*.

To retrieve **the most up-to-date** sequence, use the accession number without version: the most up-to-date sequence will be retrieved automatically.

Let’s look at the GenBank record for human alcohol dehydrogenase VI https://www.ncbi.nlm.nih.gov/nuccore/AH001409 ...

*The GI (GenInfo Identifier) system was deprecated as of 2016 – use Accession.Version only to retrieve a specific sequence – but you will still see GIs in GenBank records!

GenBank Flatfile Format (GBFF)

Homo sapiens alcohol dehydrogenase 6 (ADH6) gene, complete cds

GenBank: AH001409.2

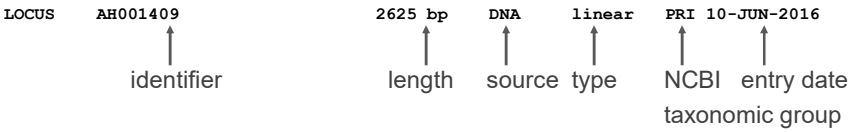
FASTA Graphics

Go to:

```
LOCUS       AH001409               2625 bp    DNA     linear   PRI 10-JUN-2016
DEFINITION  Homo sapiens alcohol dehydrogenase 6 (ADH6) gene, complete cds.
ACCESSION   AH001409 M68895 M84402 M84403 M84404 M84405 M84406 M84407 M84408
            M84409
VERSION     AH001409.2
KEYWORDS    .
```

The GenBank flatfile format (GBFF) is one of the most commonly used formats used for nucleotide sequences. It contains all of the information associated with the sequence, as well as the sequence itself.

The GBFF has 3 parts: the **header**, the **features**, and the **sequence** itself.



GenBank Flatfile Format – Header

```
DEFINITION  Homo sapiens alcohol dehydrogenase 6 (ADH6) gene, complete cds.
ACCESSION   AH001409 M68895 M84402 M84403 M84404 M84405 M84406 M84407 M84408 M84409
VERSION     AH001409.2
KEYWORDS    .
```

- DEFINITION: The biology of the molecule in a sentence.
- ACCESSION: Code(s)
- VERSION: Number; GI number found on this line too.
- KEYWORDS: Keywords as defined by the submitters…but: free-text.

GenBank Flatfile Format – Header, cont.

```
SOURCE      Homo sapiens.
ORGANISM    Homo sapiens
            Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
            Mammalia; Eutheria; Primates; Catarrhini; Hominidae; Homo.
REFERENCE   1  (bases 1 to 1925)
AUTHORS     Yasunami,M., Chen,C.S. and Yoshida,A.
TITLE       A human alcohol dehydrogenase gene (ADH6) encoding an additional
            class of isozyme
JOURNAL     Proc. Natl. Acad. Sci. U.S.A. 88 (17), 7610-7614 (1991)
PUBMED      1881901
COMMENT     On or before Jun 10, 2016 this sequence version replaced gi:178137,
            gi:178138, gi:178139, gi:178140, gi:178141, gi:178142, gi:178143,
            gi:178144, gi:178145.
```

- SOURCE: Contains organism name
- ORGANISM: Contains complete taxonomic information from the NCBI taxonomy server.
- REFERENCE: Details on a publication about the sequence.
- COMMENT: Contains misc. information and revision details.

GenBank Flatfile Format – Features

```
FEATURES             Location/Qualifiers
     source          1..2625
                     /organism="Homo sapiens"
                     /mol_type="genomic DNA"
                     /db_xref="taxon:9606"
                     /sex="male"
     regulatory      34..48
                     /regulatory_class="other"
                     /tissue_type="liver"
     mRNA            join(287..396,522..623,749..890,1016..1103,1229..1445,
                     1571..1831,1957..2092,2218..2559)
                     /gene="ADH6"
                     /product="alcohol dehydrogenase 6"
     exon            287..396
                     /gene="ADH6"
                     /number=1
```

- A direct representation of the biological information in the record.
- The Source Feature must be present in all GenBank records, and contains information as to where the molecule comes from /organism = “Homo sapiens”, and, potentially, map, chromosome and tissue type information.
- The exon feature tells one that the sequence from 287..396 comprises an exon, the first one.

GenBank Flatfile Format – Features, cont.

In some records the CDS (coding sequence) feature is present:

```
FEATURES             Location/Qualifiers
                     …
     CDS             join(379..396,522..623,749..890,1016..1103,1229..1445,
                     1571..1831,1957..2092,2218..2360)
                     /gene="ADH6"
                     /codon_start=1
                     /product="alcohol dehydrogenase 6"
                     /protein_id="AAA35509.1"
                     /translation="MSTTGQVIRCKAAILWKPGAPFSIEEVEVAPPKAKEVRIKVVAT
                     GLCGTEMKVLGSKHLDLLYPTILGHEGAGIVESIGEGVSTVKPGDKVITLFLPQCGEC
                     TSCLNSEGNFCIQFKQSKTQLMSDGTSRFTCKGKSIYHFGNTSTFCEYTVIKEISVAK
                     IDAVAPLEKVCLISCGFSTGFGAAINTAKVTPGSTCAVFGLGGVGLSVVMGCKAAGAA
                     RIIGVDVNKEKFKKAQELGATECLNPQDLKKPIQEVLFDMTDAGIDFCFEAIGNLDVL
                     AAALASCNESYGVCVVVGVLPASVQLKISGQLFFSGRSLKGSVFGGWKSRQHIPKLVA
                     DYMAEKLNLDPLITHTLNLDKINEAVELMKTGKW"
```

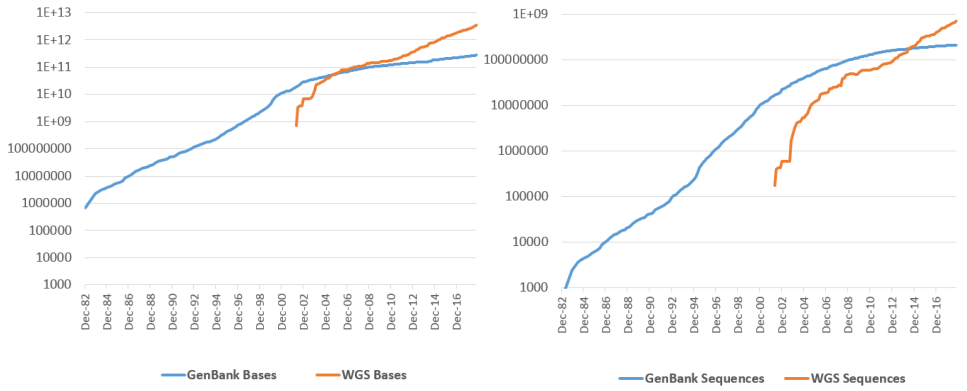

GenBank Flatfile Format – Sequence

The last part of the GenBank flat file record is the sequence itself:

```
ORIGIN
        1 tgtattttga aaacaacaga aaagaaatac ttttgtacac tctgttagaa attttaagtt
       61 tggacattta aaagtccaaa tttaaaactc aaaaaaatgg ataataagag ggacctgttt
      121 gattaaggga gaaaaaaata gtttgcattt tcacctttg gctctttcac tgagatgagc
      181 ctatttcaga ttacacttag gaacttccat caagcacggg agagcctact tttcctgttt
      241 aataattacc agactacaga gaaggtcgga ccagccttct gatctacagt cgcctgtgta
      301 cctttgtact ttctacagtg aaagttgcta caggatctcc ctttctcaat aaattcatct
      361 gcggtggaga aaatcagcat gagtactaca ggccaagtag gtgcagtat
          …
     2461 actgataatt gaagaggctt tcaggaattt gtaaagcatc tccttcccct ctgcattttg
     2521 ttttatttct agctaataaa atacataatc ctgaaagtat ttaagtgttc acctaccgtt
     2581 acttttgcca attagcattg tatttccaat atggattttt ttttt
//
```

Nucleotide Databases – Growth of GenBank

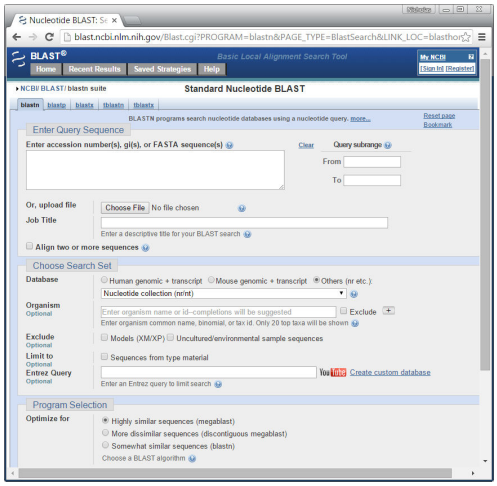
from http://www.ncbi.nlm.nih.gov/genbank/statistics



Searching GenBank + other sequence DBs

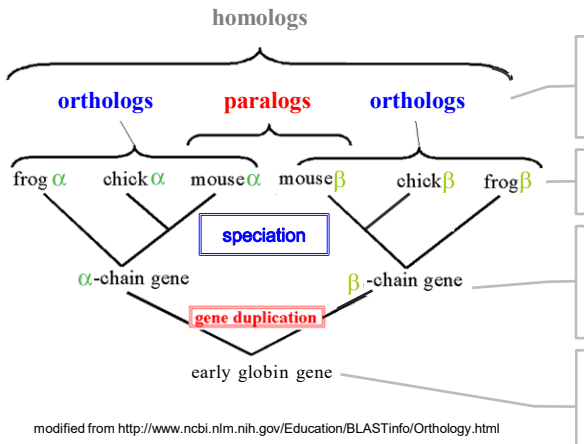
→by keyword

→by sequence similarity, using BLAST* (http://www.ncbi.nlm.nih.gov/BLAST/)



*Google and other search engines don't handle sequence searches well: they can't put in gaps to identify partial matches to similar sequences, and they don't know which amino acids have similar properties!

Definitions



modified from http://www.ncbi.nlm.nlm.nih.gov/Education/BLASTinfo/Orthology.html

Orthologs are related by a speciation event, paralogs by duplication

Present day species' genes

Gene duplication results in two copies in common ancestor

Start: just one copy of globin in ancestral organism

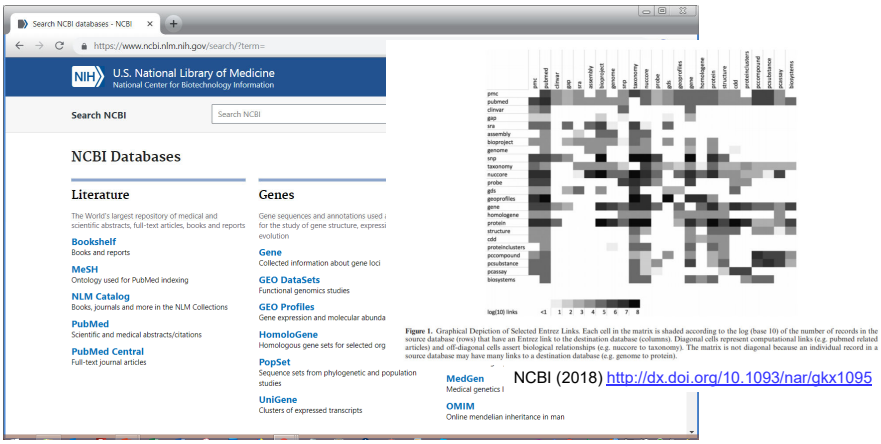
Searching across DBs: the NCBI Search tool

Several publically-available tools are available for querying across databases. One is provided by the NCBI and is called Search (formerly Entrez/GQuery) (https://www.ncbi.nlm.nih.gov/search/). NCBI Search essentially provides links between many of the databases at NCBI.

We'll go through an example using NCBI Search...

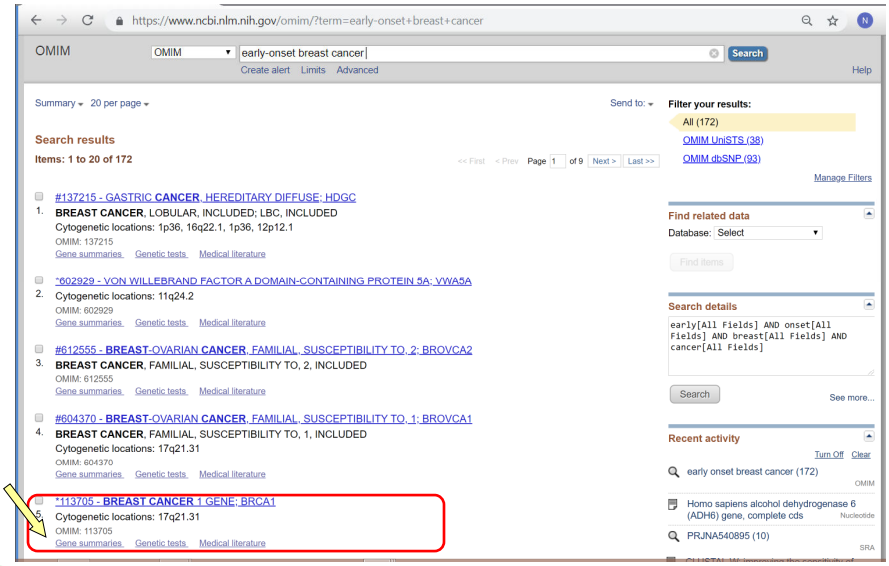
Sample Problem

Identify the SNPs which potentially cause early onset breast cancer, and design oligos to PCR them in samples of human genomic DNA for sequencing. Use the OMIM "function" of NCBI Search. OMIM has links to everything that is known about a given disease across the various databases at NCBI.

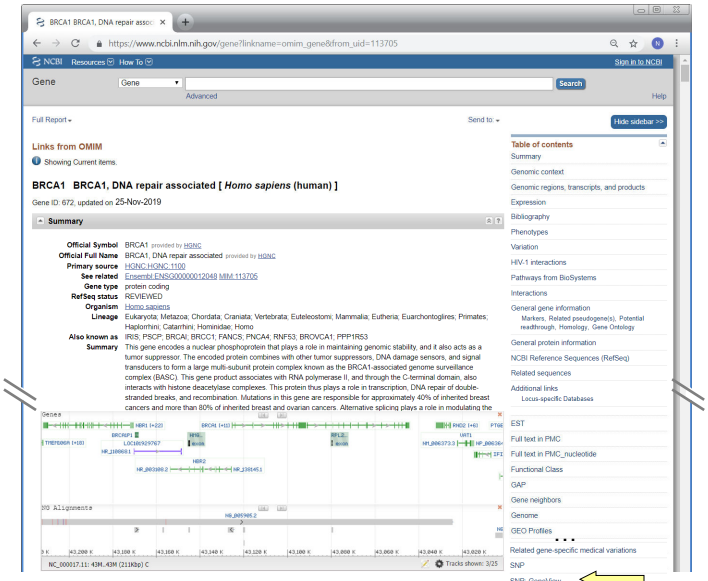


NCBI (2018) http://dx.doi.org/10.1093/nar/gkx1095

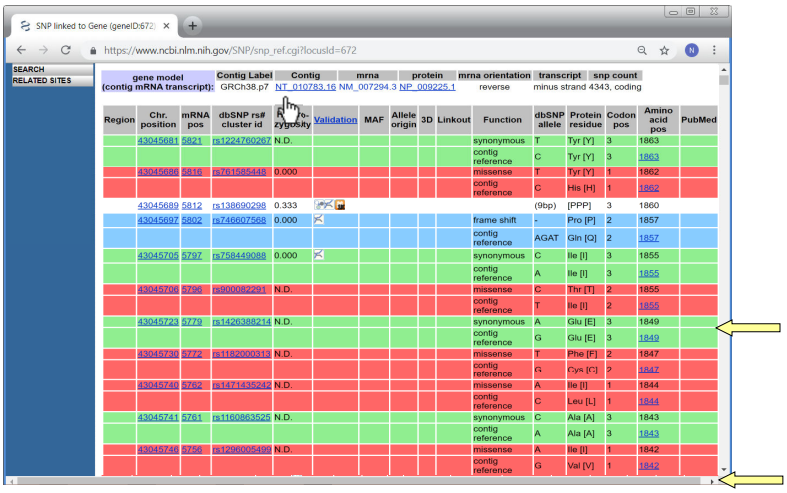
Sample Problem [2]



Sample Problem [4]

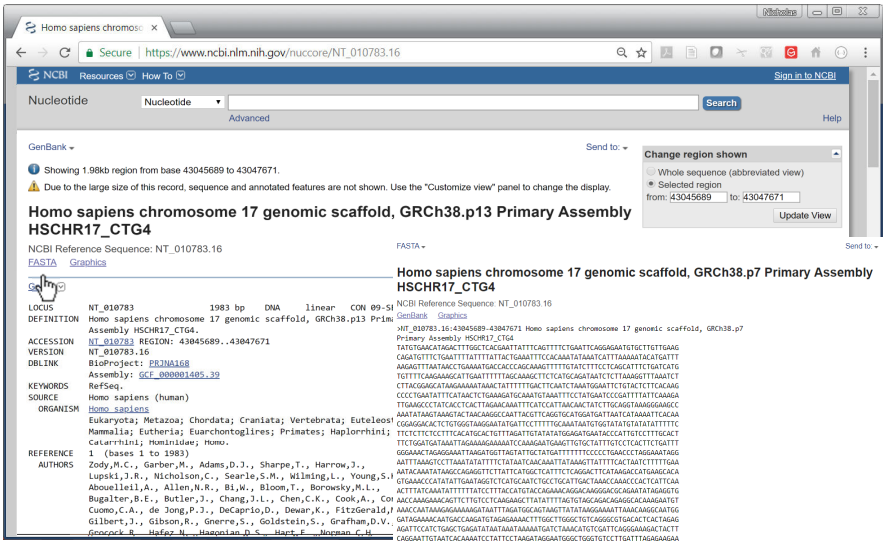


Sample Problem [5]



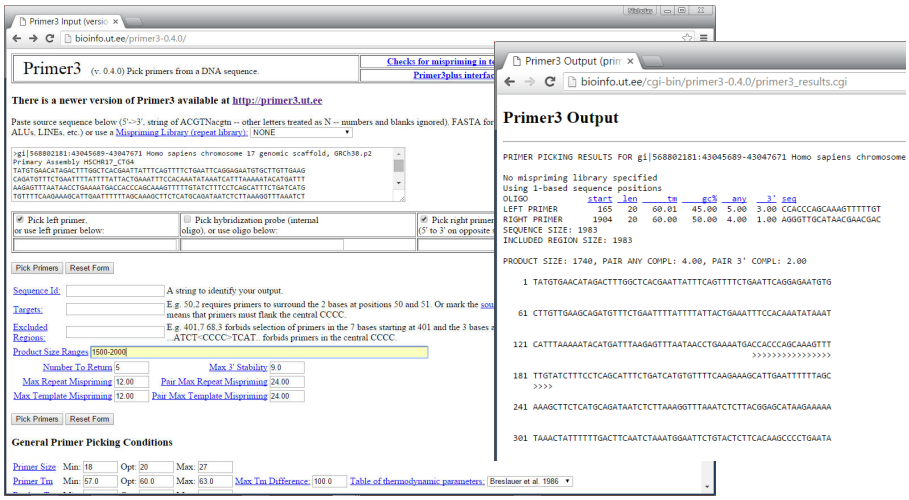
Choose start and stop in Contig (next slide) – don't need whole chromosome!

Sample Problem [6]



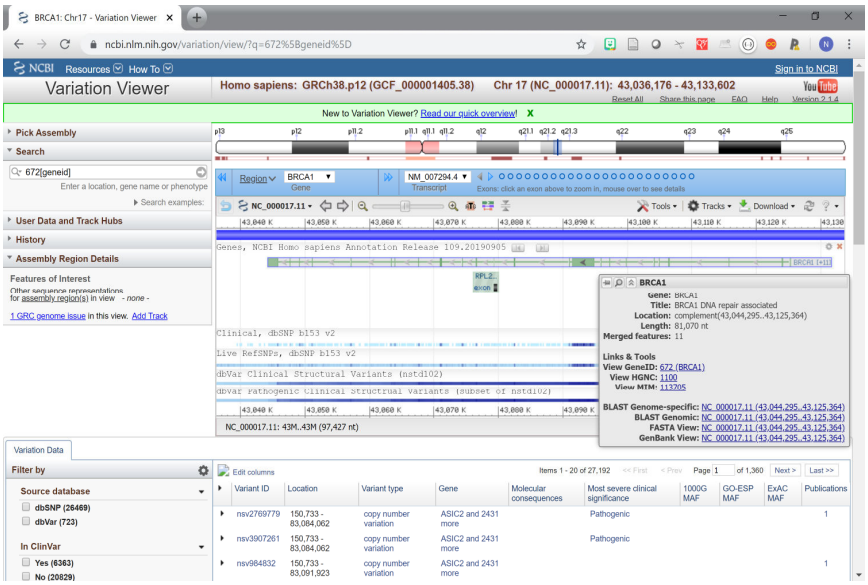
Primer3 can then be used to design PCR primers

Use it at http://frodo.wi.mit.edu/



Steve Rozen and Helen J. Skaletsky (2000), in: Krawetz S, Misener S (eds) Bioinformatics Methods and Protocols: Methods in Molecular Biology. Humana Press, Totowa, NJ, pp 365-386

Sample Problem [5 – new Variation Viewer]



Which Database for What?