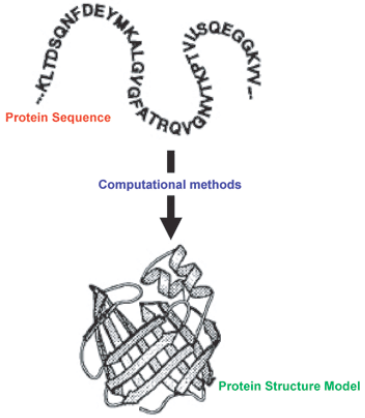
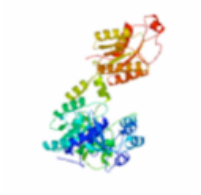
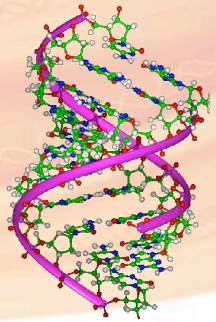




Bioinformatics Computational Methods 1 - BIOL 6308



October 15th 2013

http://155.33.203.128/cleslin/home/teaching6308F2013.php

Last Time

- Why Study Genomes
- Genome Organization
  - Genes in Bacteria
  - Genes in Eukaryotes
  - Control Regions
  - Exon/Intron Structure
- Sensitivity and Specificity
- Locating Genes in Genomes
  - Use coding regions from other organisms
  - Identify genes from the properties of the DNA sequences themselves
    - *Ab initio* methods
    - GENESCAN
    - Evaluation of Methods
    - Sequence Coding Statistics
  - Comparative genome approach

Log-Likelihood Ratio - *LLR*

- Let ***F(c)*** be the frequency (probability) of codon ***c*** in the genes of the species under consideration (use published data)
- Then, given a sequence of codons , $C = C_1 C_2 \cdots C_m$ and assuming independence between adjacent codons

$$P(C) = F(C_1)F(C_2)\cdots F(C_m)$$

is the probability of finding the sequence of codons C knowing that *C* codes for a protein

- For instance, if ***S*** is the sequence ***S*** = AGGACG, when read in frame 1, it results in the sequence $C_1^1 = \mathtt{AGG}\ \ C_2^1 = \mathtt{ACG}$

- Then $P^1(S) = P(C^1) = F(\mathtt{AGG})F(\mathtt{ACG})$
- Substituting the appropriate values from published data, we obtain

$$P^1(S) = P(C^1) = 0.022 \times 0.038 = 0.000836$$

Log-Likelihood Ratio - *LLR*

- On the other hand, let $F_0(c)$ be the frequency of codon c in a non-coding sequence

$$P_0(S) = P_0(C) = F_0(C_1)F_0(C_2)\cdots F_0(C_m)$$

is the probability of finding the sequence S if C is non-coding

- Assuming the random model of coding DNA is $F_0(c) = 1/64 = 0.0156$ for all codons
- P_0 for the above sequence of codons C would be:

$$P_0(C) = 0.0156 \times 0.0156 = 0.000244$$

- The LLR ratios for S codling in frame 1, LP^1 is

$$LP^1(S) = \log(0.000836/0.000244) = \log(3.43) = 0.53$$

Remember, LLR used throughout bioinformatics

Why Study Genomes?

- Understand biological processes
- Understand pathological processes
- Diagnose, prevent, and cure diseases

- Outline
  - Genome Organization
  - Locating Genes in Genomes
  - **What About Other Genomic Features**
  - **And now the Genomes…**

What About Other Genomic Features

Other Genomic Features

- Other than protein coding genes, what is there?
  - Genes for noncoding RNAs (rRNA, tRNA, miRNAs, etc.)
  - Structural sequences (scaffold attachment regions)
  - **Regulatory proteins and regions**
  - Non-functional
- We can begin to annotate regulatory sequences such as transcription factor (**TF**) binding sites and cis-regulatory modules
  - Gene regulation
  - Regulatory proteins

Combinatorial Gene Regulation

- A microarray experiment showed that when gene **X** is knocked out, **20** other genes are not expressed
  - **How can one gene have such drastic effects?**

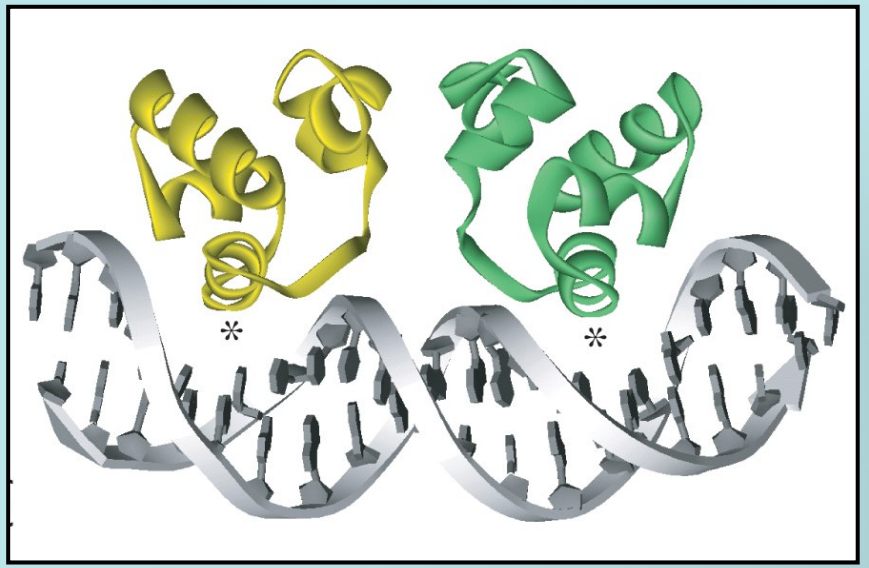
Regulatory Proteins

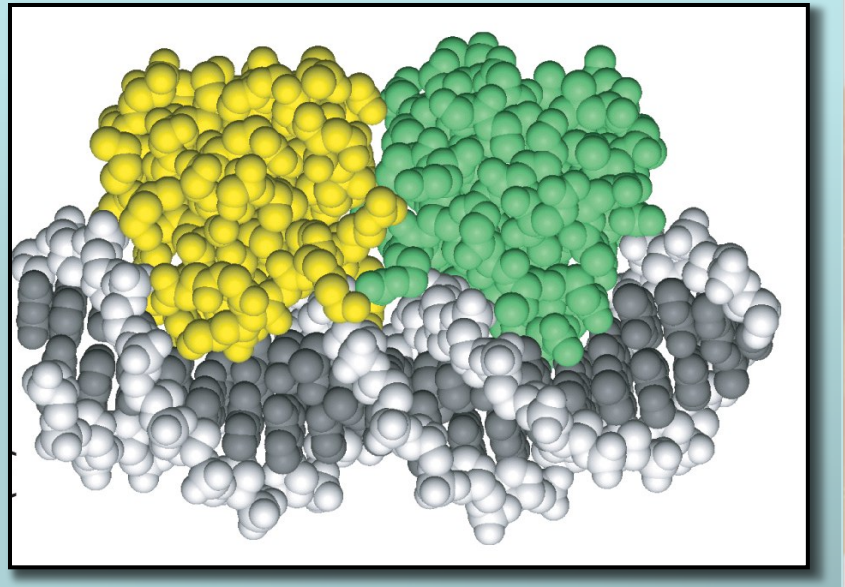
- Gene X encodes a regulatory protein:
  - a.k.a. a TF
- The 20 unexpressed genes rely on gene X's **TF** to **induce** transcription
- A single TF may regulate multiple genes

Regulatory Regions

- Every gene contains a regulatory region (RR) typically stretching 100-1000 bp upstream of the transcriptional start site
- Located within the RR are:
  - Transcription Factor Binding Sites (**TFBS**)
    - known as motifs - specific for a given transcription factor
  - **TFBS** can be located anywhere within RR
  - **TFBS** may **vary slightly** across **different** RR since **non-essential bases can mutate**
- **TFs** influence gene expression by binding to a specific location in the respective gene's regulatory region - **TFBS**

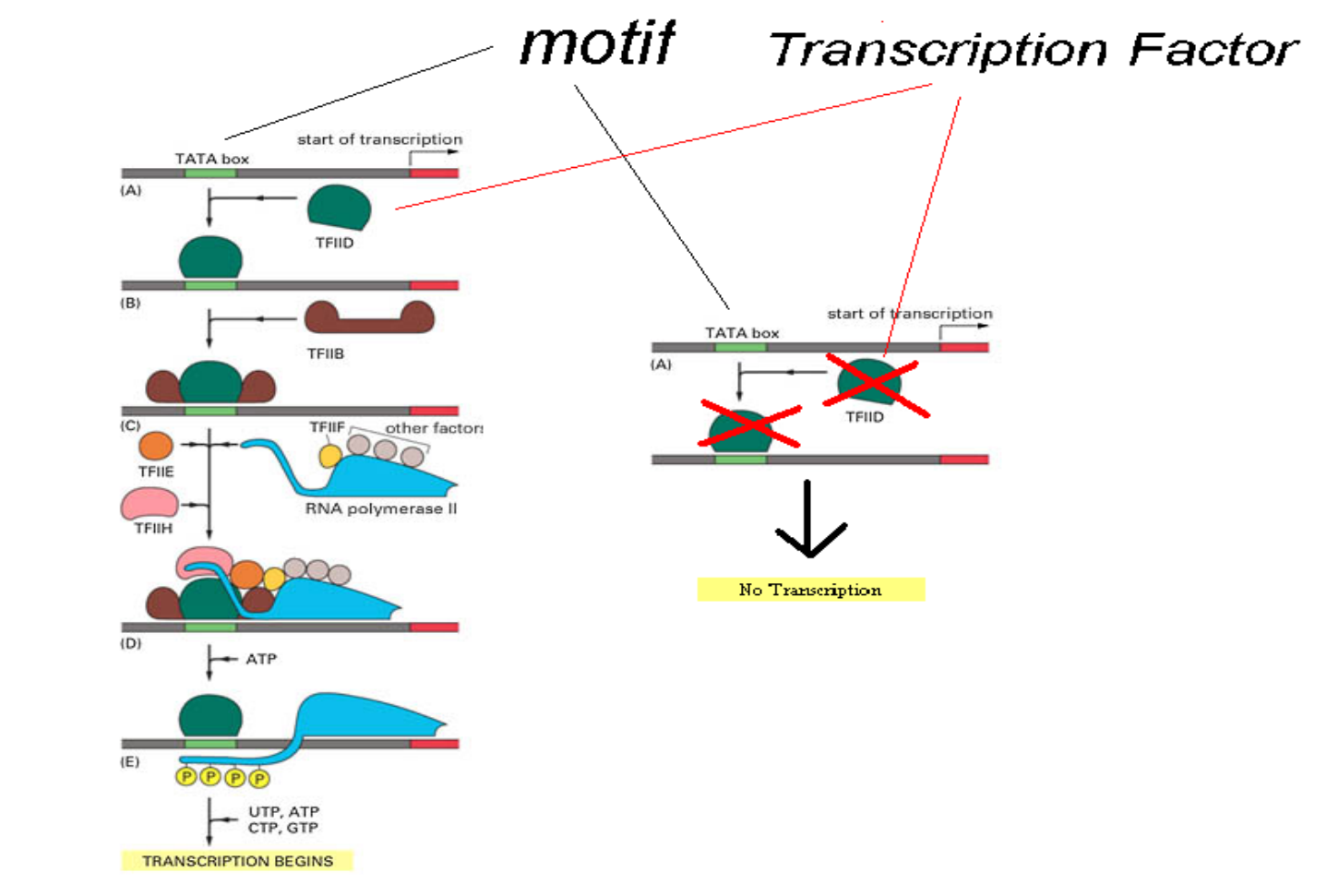
Control of Gene Expression—Transcription Factors



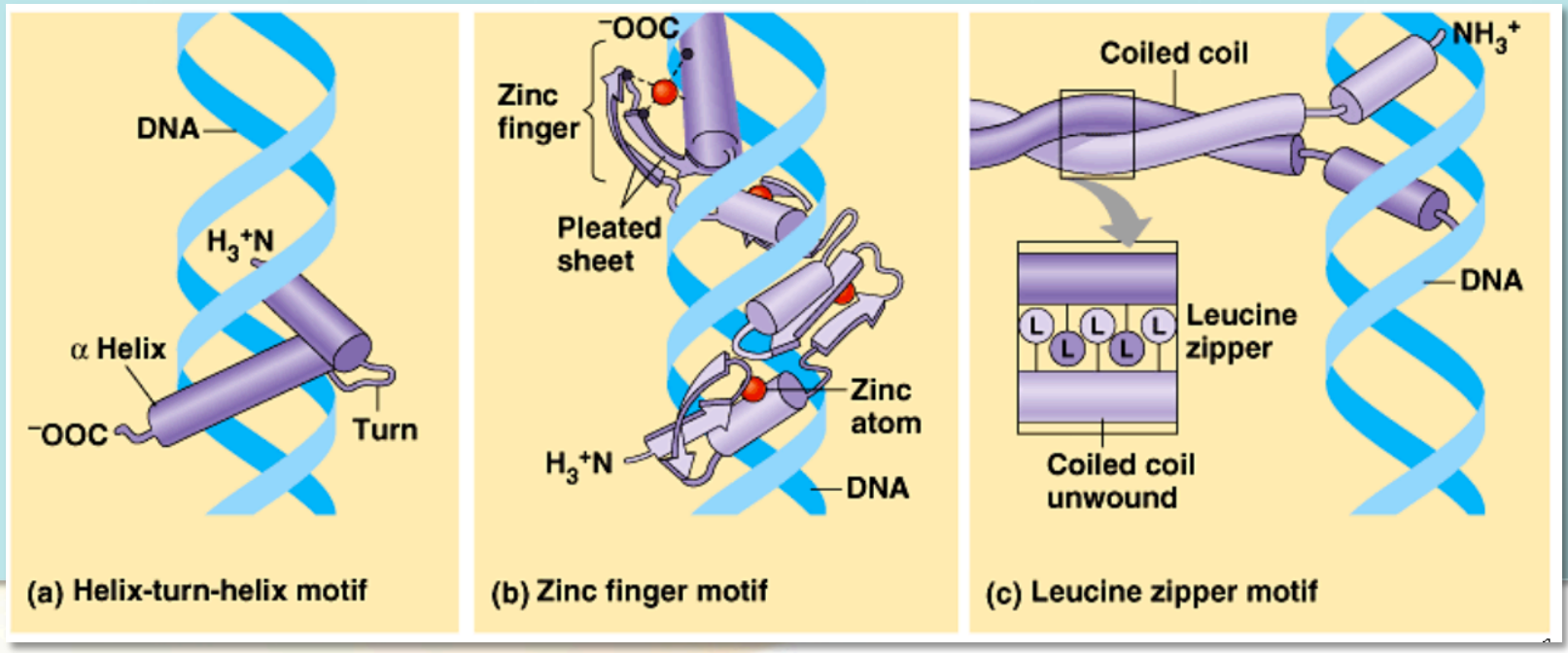


TFs are proteins that bind to the DNA and help to control gene expression. We call the sequences which TF bind - ***TFBSs***, which are a type of ***cis*-regulatory sequence**

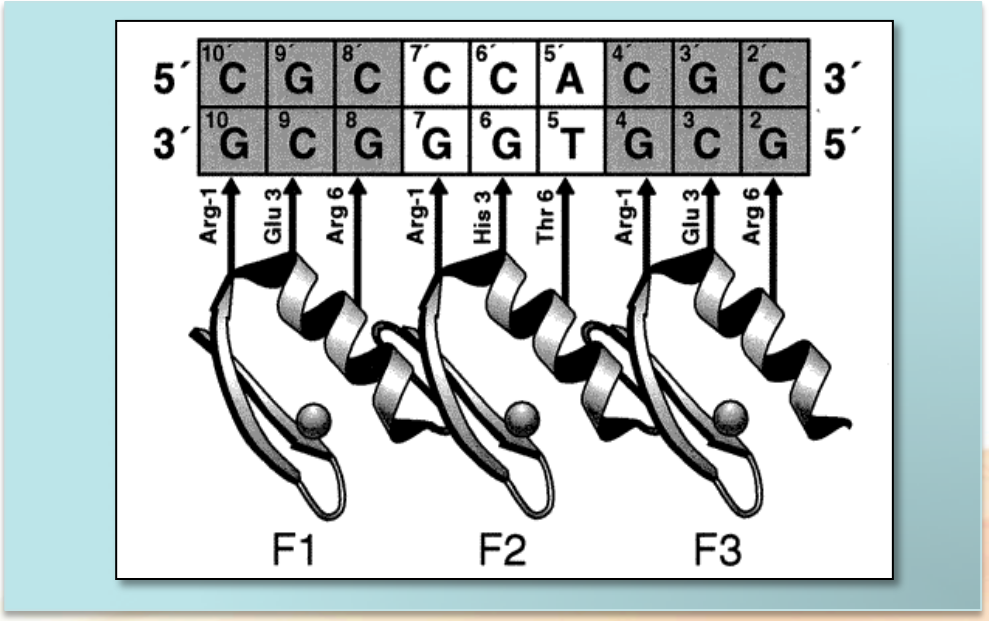
TFs and Motifs



Types of DNA Binding Domains



TFs Bind to Specific DNA Sequences

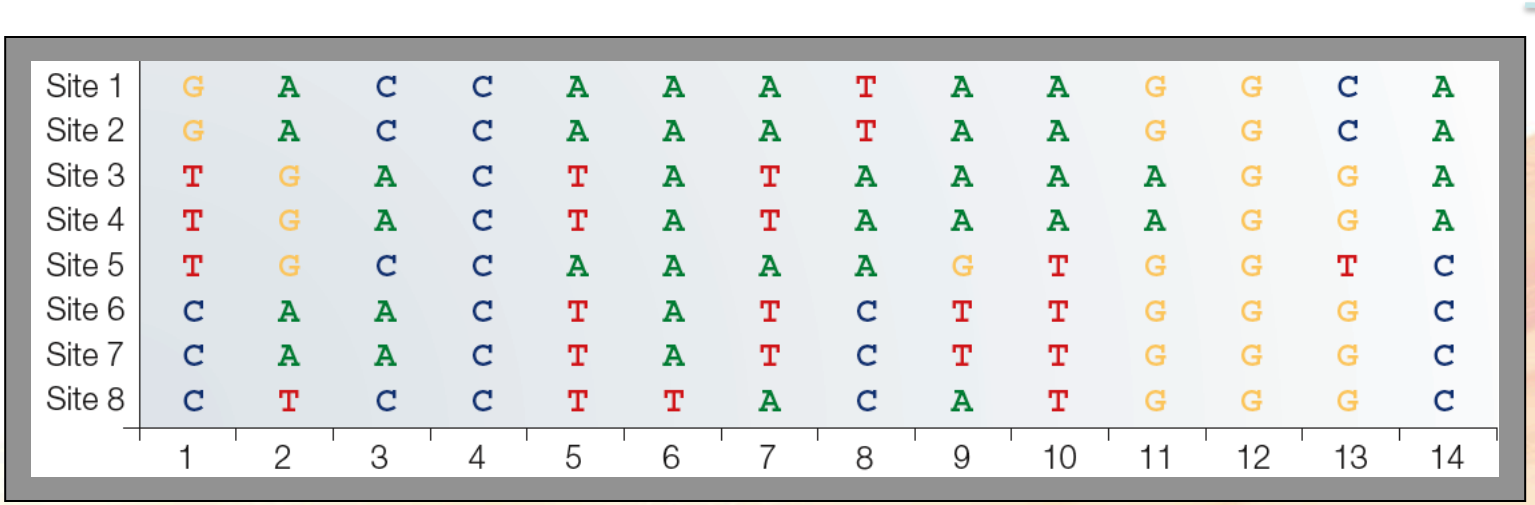


Binding sites are **usually** first determined empirically

Once we know the **binding site**, we can search the **genome** to find *all* of the (predicted) binding sites

Control of Gene Expression - TFs

- Most transcription factors can bind to a range of similar sequences
- We call this a binding "motif"



a range of similar sequences

Wasserman, W. W. and A. Sandelin (2004). *Nat Rev Genet* **5**(4): 276-287.

- We can represent these in either of two ways:
  - as a ***consensus sequence***
  - or as a ***position weight matrix (PWM)***

Binding Site (motif) Representations

- **7 characterized binding sites for a certain transcription factor:**

TCCGGAAGC
TCCGGATGC
TCCGGATCT
CATGGATGC
CCAGGAAGT
GGTGGATGC
ACCGGATGC

- **consensus sequence (next slide)**

[T/C]C[C/T]GGAAGC

- **Frequency matrix** and its graphical depiction, **a *sequence logo***

A 111007200
T 302000502
G 110770060
C 254000015

Binding Site (motif) Representation

- **Consensus sequence is one-line description of TFBS**
  - Based on column-by-column alignment of individual known binding sites
    A Common rule is:

> A single base shown if it occurs in **more than half the sites** and at **least twice as often** as the **second most frequent base**. Otherwise, a **double degenerate symbol** (e.g., G/C= S) is used if two bases occur in more than 75% of the sites, or a triple degenerate symbol when one base does not occur at all.

- Think of consensus as an "ancestor" motif, from which mutated motifs emerged
- The *distance* between a real motif and the consensus sequence is generally less than that for two real motifs
  - Distance?

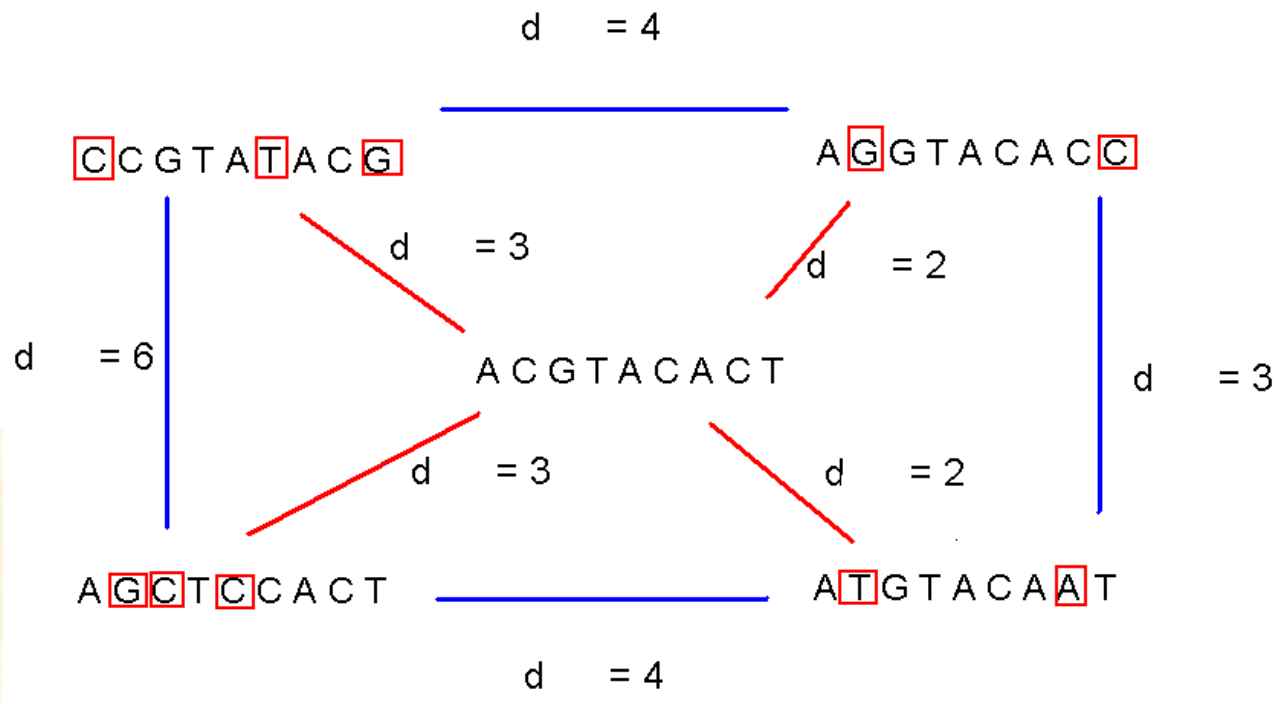
IUPAC Ambiguity Code

- Degenerate base symbols in biochemistry are an IUPAC representation for a position on a DNA sequence that can have multiple possible alternatives

IUPAC Code	Mnemonic	Meaning	Complement
A	Adenine	A	T
C	Cytosine	C	G
G	Guanine	G	C
T/U	Thymidine	T	A
K	Keto	G or T	M
M	Amino	A or C	K
S	Strong	C or G	S
W	Weak	A or T	W
R	Purine	A or G	Y
Y	Pyrimidine	C or T	R
B	not A	C or G or T	V
D	not C	A or G or T	H
H	not G	A or C or T	D
V	not T and not U	A or C or G	B
N	any	G or A or T or C	N

Consensus

The *distance* between a real motif and the consensus sequence is generally less than that for two real motifs



Much more on distance in upcoming lectures

Finding Binding Sites in the Genome

$^{T}_{C}C^{C}_{T}GGATGC$

- Consensus sequences **make searching straightforward**
- Simple text search that can even be done using a word processor, or Perl program:

```
while(<SEQUENCE>){
   if ($_ =~ /[T|C]C[T|C]GGATGC/){
      do something;
   }
}
```

- All positions in the motif are treated the same

Issues with Finding Binding Sites in the Genome

- Use caution
  - Just because a sequence in genome is **reasonable match** to a known TFBS
    - Doesn't necessarily mean that the TF is binding there *in vivo*
    - Why is that exactly?
      - By crude calculation:

The probability of finding a 7 bp motif is 4^7 = 1/16,384
i.e., expect only about 1 motif every 16 kb.

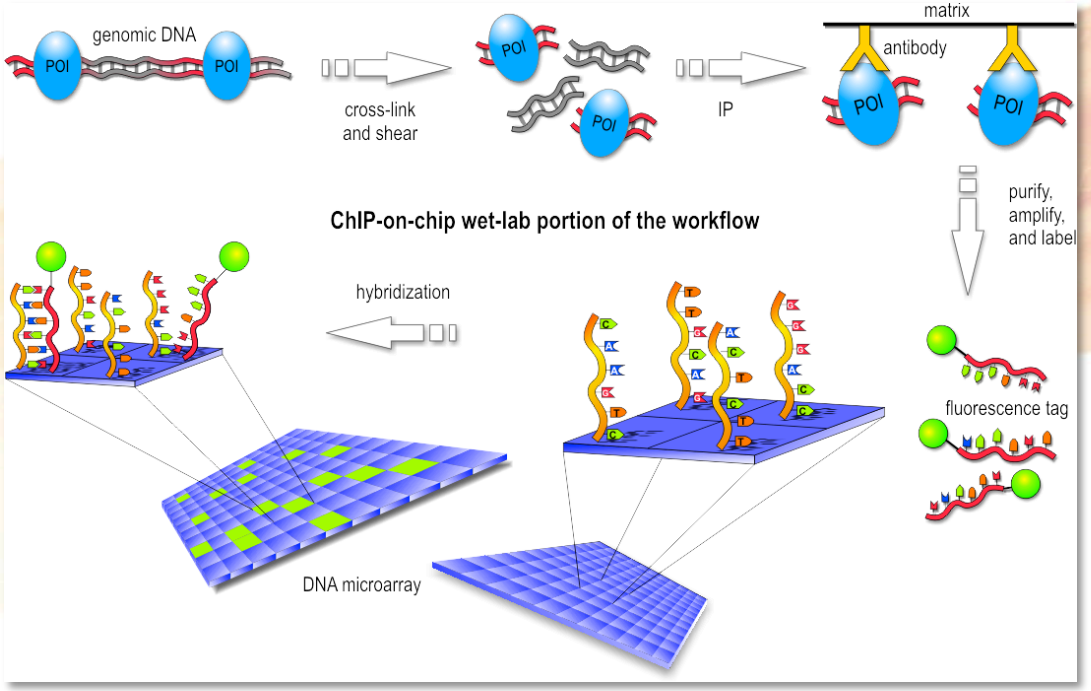
So in human genome, this sequence should be present over **183,000** times!
(~9x per gene!) Even in a 10 Mb genome, the sequence would occur over 600 times.

This calculation does not even take into **account motif degeneracy**!

- Need to consider additional factors in deciding what predicted binding sites are important—such as how regulatory regions are organized
- **We need a more sophisticated model**

Empirical Methods for TFBS

- Empirical methods, such as ChIP-chip are a good alternative for looking at *in vivo* binding
- Allows for identification of the cistrome - Sum of binding sites
- Bioinformatics methods combined with this method – very effective determine the transcription factor binding motifs



poi = protein of interest

Genome Annotation - Transcription Factors

- Because difficulty in accurately predicting true, functional TFBSs
  - **Most genome annotation** focuses on empirically determined sites
- Several databases curate these data:
  - Tracks displaying these data can be found in the UCSC Genome Browser
    - *Open Regulatory Annotation database (ORegAnno)*
    - *Regulatory Element Database for Drosophila (REDfly)*
    - Also curate *cis*-regulatory module sequences, which at present can only reliably be determined by empirical methods
  - Jaspar database - experimentally defined transcription factor binding sites for eukaryotes
  - Plant TFBS Database - arising from efforts to identify and catalogue all *Plant* genes involved in transcriptional control
  - AtTFDB – specialized for Arabidopsis

Genome Annotation - Transcription Factors Links Previous Page

- http://genome.ucsc.edu
- http://www.oreganno.org/oregano/Index.jsp
- http://redfly.ccr.buffalo.edu/
- http://jaspar.genereg.net/
- http://plntfdb.bio.uni-potsdam.de/v3.0/
- http://arabidopsis.med.ohio-state.edu/AtTFDB/

And Now the Genomes…

What's in a Genome?

- Genes (i.e., protein coding)
- But. . . only <2% of the human genome encodes proteins
- Other than protein coding genes, what is there?
  - Genes for noncoding RNAs (rRNA, tRNA, miRNAs, etc.)
  - Structural sequences (scaffold attachment regions)
  - **regulatory proteins and regions**
  - Non-functional
- It's still uncertain/controversial how much of the genome is composed of any of these classes
- The answers will come from experimentation and bioinformatics

Genomes

- Typical bacterium comes as a single DNA molecule of about **5 million** characters
  - ~Fairly large book
  - Extended ~ about **2 mm** long
    - It has to fit into a cell of diameter of about 0.001 mm
- DNA of higher organisms is organized into chromosomes
  - example: Humans contain 23 chromosome pairs
- The total amount of genetic information per cell:
  - **The sequence of nucleotides of DNA, nearly constant for all members of a species**
  - **But varies widely between species**

Genome Sequencing

6/25/04	(7/25/05)	(7/20/07)	**6/29/08**	***10/07/2011***	
1128	**(1496)**	**(2680)**	**3825**	***11509***	**genome projects:**
199	**(274)**	**(579)**	**827**	***2941***	complete (includes 28 (36) (49) **94** ***166*** euk's)
508	**(728)**	**(1285)**	**1932**	***2673***	prokaryotic genomes in progress
421	**(494)**	**(721)**	**936**	***753***	eukaryotic genomes in progress

http://www.genomesonline.org/cgi-bin/GOLD/index.cgi

Go and look how the numbers are now in 2013!

Genome Prokaryotes

- Genetic material of most prokaryotic cells takes form of a large single circular piece of double-stranded DNA
  - Typically <= 5 Mb long
  - Cells may contain plasmids
- In many prokaryotic genomes protein-coding regions partially organized into ***operons***
  - Tandem genes transcribed into a single mRNA molecule
  - Under common transcriptional control

Genome Prokaryotes - Continued

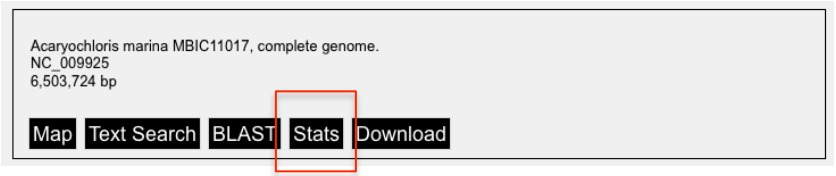
- **FYI - Bacteria**
  - Genes of many operons code for proteins with related functions
  - i.e. successive genes in the trp operon of E. coli code for proteins that catalyse successive steps in the biosynthesis of tryptophan
- **Archaea**
  - a metabolic relationship between genes in operons is **less frequently observed**
- Typical prokaryotic genome contains only a **relatively small amount of non-coding DNA** (in comparison with eukarya)
  - Distributed throughout the sequence
  - i.e. E. coli, only ~11% of the DNA is non-coding

The Genome of the Bacterium E. coli

- Strain K-12 - workhorse of molecular biology
- Genome of strain MG1655, published 1997, contains **4,639,221 bp** in a single **circular DNA molecule**, no plasmids
- ~89% of the sequence codes for proteins or structural RNAs
- **An inventory reveals:**
  - **4284** protein-coding genes
  - **122** structural RNA genes
  - non-coding repeat sequences
  - regulatory elements
  - transcription/translation guides
  - transposase
  - prophage remnants
  - insertion sequence elements
  - patches of unusual composition, likely to be foreign elements introduced by horizontal transfer

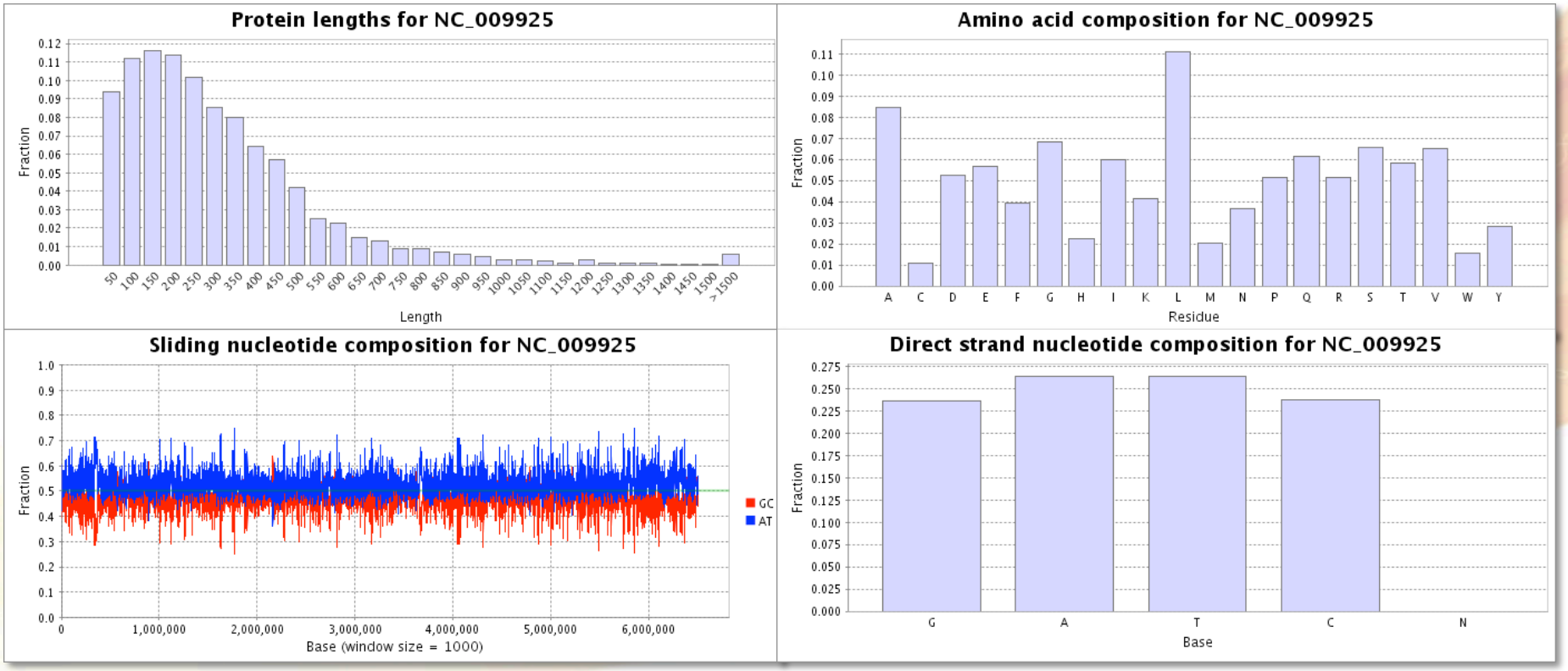
Maps of Bacterial Genomes

- BacMap - http://wishart.biology.ualberta.ca/BacMap/index.html
  - Interactive visual database containing hundreds of fully labeled and searchable maps of bacterial genomes
- Maps for 2023 bacterial chromosomes
  - Supports zooming and rotation
  - Hyperlinked to detailed textual annotations
  - Can be explored manually, or with the help of BacMap's built in text search and BLAST search
- A written synopsis each bacterial species is provided
- Charts illustrating the proteomic and genomic characteristics of each chromosome are available
- **Flat file versions** of the BacMap gene annotations, gene sequences and protein sequences can be downloaded



Acaryochloris marina MBIC11017

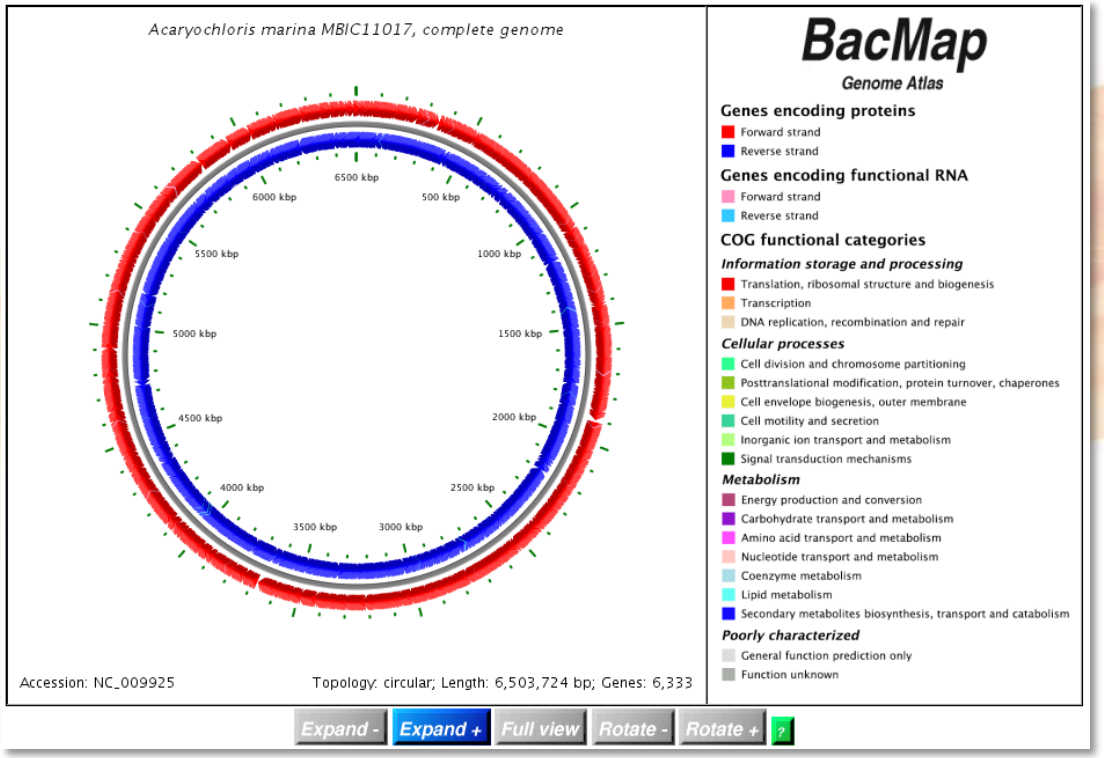
- High level overview of DNA and protein composition of the genome



http://wishart.biology.ualberta.ca/BacMap/cgi/getGraphs.cgi?accession=NC_009925&ref=index.html

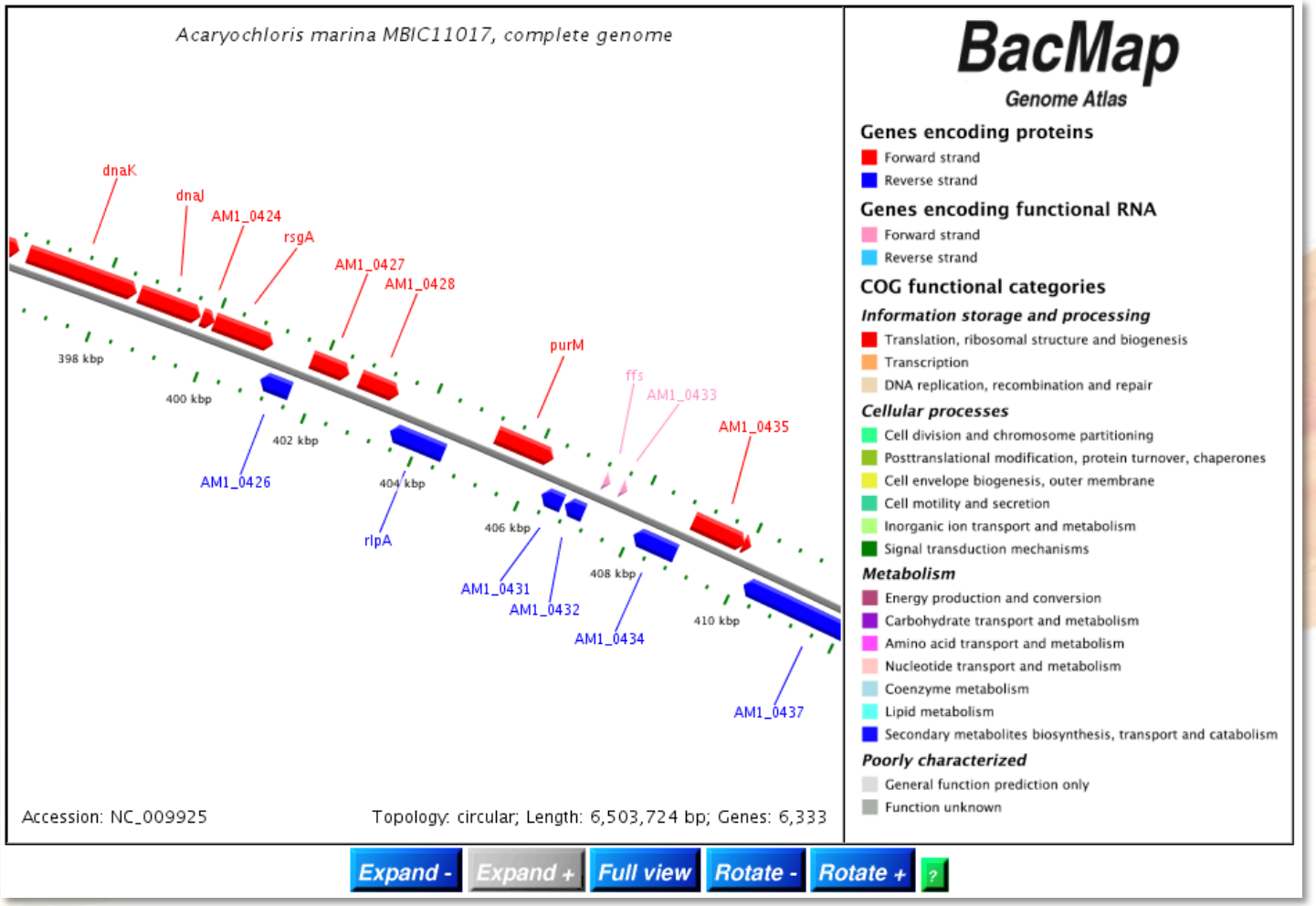
Visualizing Genomes BacMap

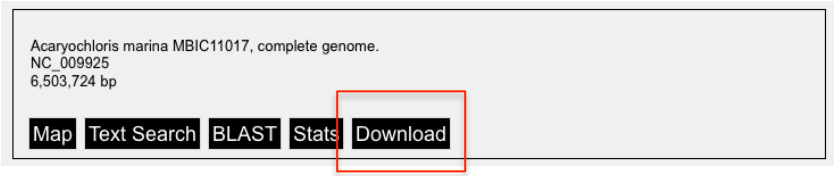
- Visualize contents of bacterial and organelle genomes as concentric circular diagrams
- Complex patterns of color coding serve as a visual 'feature table'
- http://wishart.biology.ualberta.ca/BacMap/index.html



http://wishart.biology.ualberta.ca/BacMap/cgview_linked_maps/NC_009925/index.html

How Visualize Genomes





Downloads: Genomes BacMap

- Download protein, protein encoding genes, functional RNA encoding gene sequences, Chromosome sequences

Back

Definition: Acaryochloris marina MBIC11017, complete genome.
Accession: NC_009925
Length: 6,503,724 bp

Downloads:

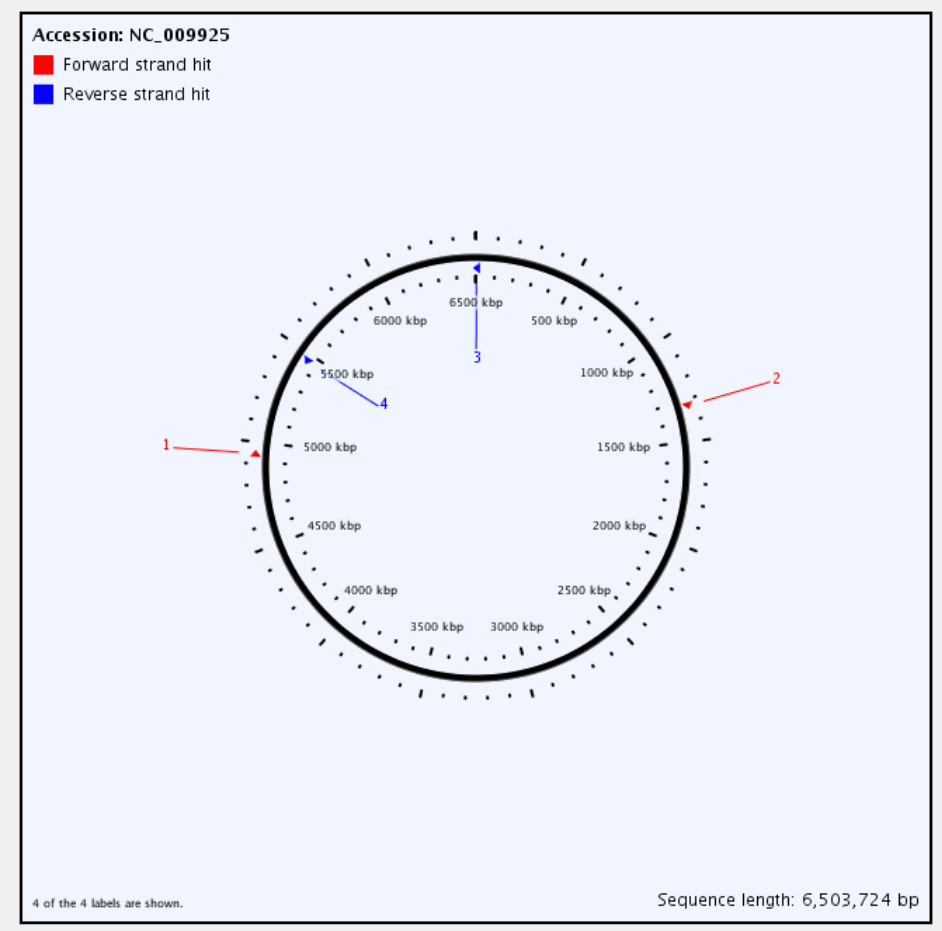
Protein sequences
Protein encoding gene sequences
Functional RNA encoding gene sequences
Chromosome sequence
GenBank record
The BacMap gene cards for NC_009925 are currently not available in tar.gz format
The BacMap gene cards for NC_009925 are currently not available in zip format

Back

http://wishart.biology.ualberta.ca/BacMap/cgi/getDownloads.cgi?accession=NC_009925&ref=index.html

Text Search: Genomes BacMap

- Did search for NUDIX hydrolase



Results 1 - 4 of 4 for **NUDIX hydrolase**

1 - **AM1_4898** NUDIX hydrolase
View map | View gene card

2 - **AM1_1353** NUDIX hydrolase
View map | View gene card

3 - **AM1_0003** NUDIX hydrolase
View map | View gene card

4 - **AM1_5401** NUDIX hydrolase
View map | View gene card

http://wishart.biology.ualberta.ca/BacMap/cgi/textSearchGenomes.cgi?accession=NC_009925&ref=index.html

Genomes Eukaryotes

- **It's rare in science to encounter a completely new world containing phenomena entirely unsuspected**
- The complexity of the eukaryotic genome is such a world

Genomes Eukaryotes: **An Inventory Reveals:**

- **Moderately repetitive DNA**
  - **Functional**
    - dispersed **gene families**
      - e.g. actin, globin
    - tandem gene family arrays
      - rRNA genes (250 copies)
      - tRNA genes (50 sites with 10–100 copies each in human)
      - histone genes in many species
  - **Without known function**
    - short interspersed elements (SINEs)
      - Alu is an example
      - 200–300 bp long
      - 100 000s of copies (300 000 Alu)
      - scattered locations (not in tandem repeats)
    - long interspersed elements (LINEs)
      - 1–5 kb long
      - 10–10 000 copies per genome –processed pseudogenes

Genomes Eukaryotes: **An Inventory Reveals:**

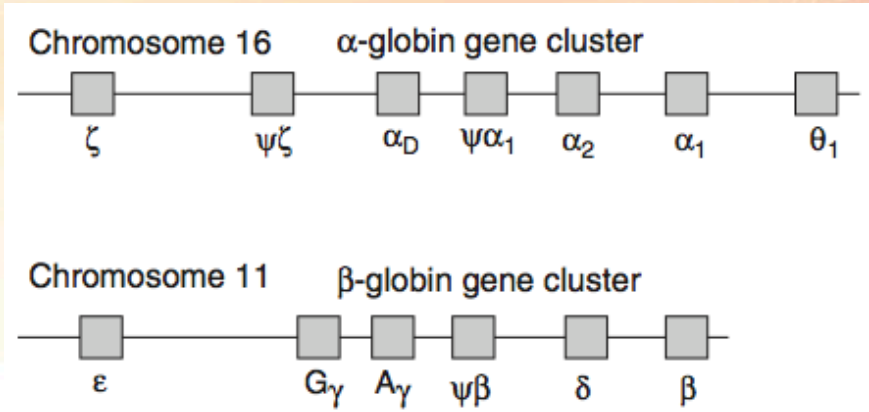
- **Highly repetitive DNA**
  - **Minisatellites**
    - composed of repeats of 14–500 bp segments
    - 1–5 kb long
    - many different ones
    - scattered throughout the genome
  - **Microsatellites**
    - composed of repeats of up to 13 bp
    - ~100s of kb long
    - ~106 copies/genome
    - most of the heterochromatin around the centromere
  - **Telomeres**
    - contain a short repeat unit
      - typically 6 bp:
        - TTAGGG in human genome
        - TTGGGG in Paramecium
        - TAGGG in trypanosomes
        - TTTAGGG in Arabidopsis)
      - 250–1000 repeats at the end of each chromosome

Gene Families

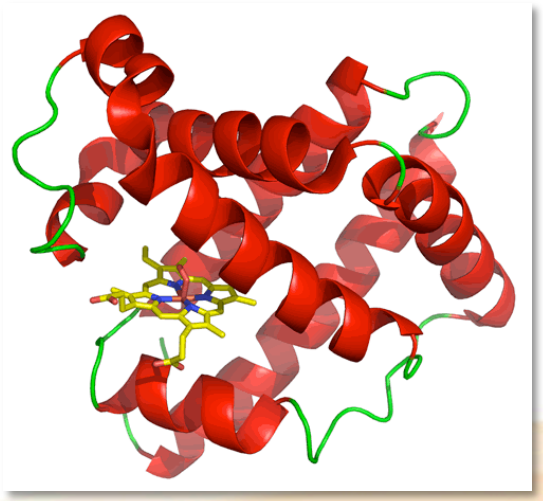
- **Gene families** within single chromosomes are common in eukaryotes
- Collection of identical or very similar genes
- Family members are:
  - **Paralogous** - related genes that have diverged to **provide separate functions in the same species**
  - **Orthologous** - homologues that **perform same function** in different species
  - **Pseudogenes** - arisen by (duplication/retrotransposition from mRNA) followed by accumulation mutations to **point of loss of function**
- For instance:
  - Human α and β globin are **paralogos**
  - Human and horse **myoglobin** are **orthologos**

The Globin Gene Cluster

- Human hemoglobin genes and pseudogenes appear clusters on chromosomes **11 and 16**
- **Adult human** synthesizes primarily **3 types of globin chains**:
  - α and β chains - assemble into hemoglobin $\alpha_2\beta_2$ **tetramers**
  - myoglobin - **monomeric** protein - muscle
- Other forms hemoglobin synthesized in embryonic and fetal stages of life
  - Encoded by different genes
- Other globins are unlinked; they arose long before this cluster diverged



Monomer vs Tetramer



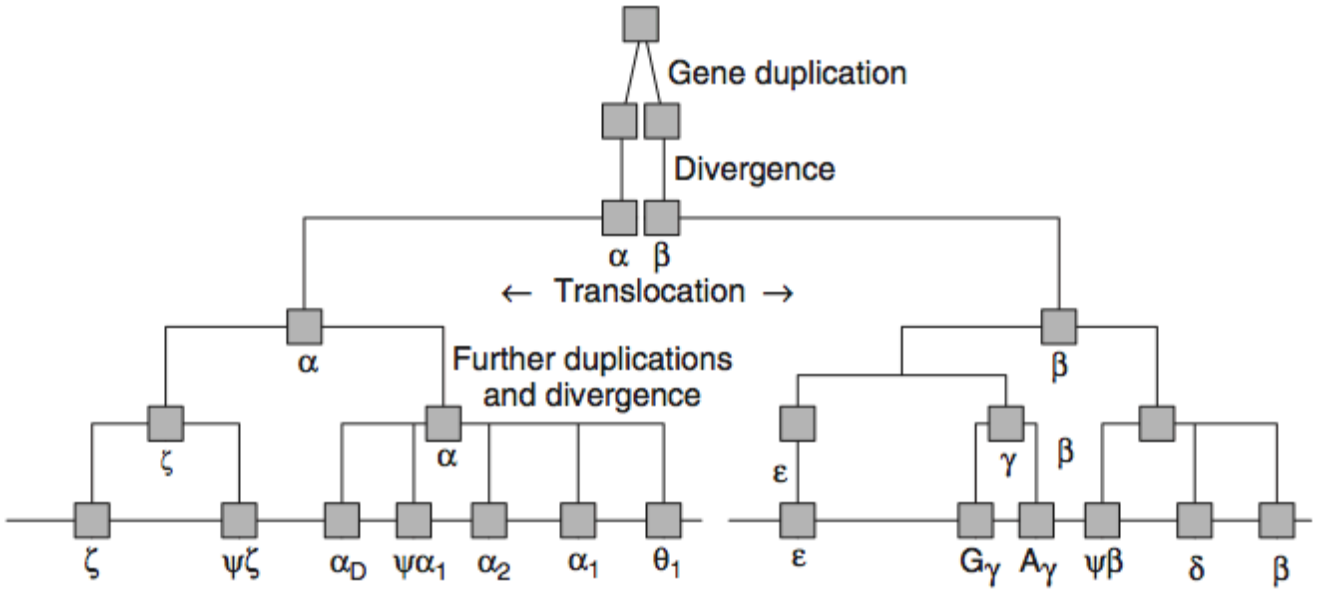
Myoglobin

Hemoglobin

http://guweb2.gonzaga.edu/faculty/cronk/biochem/m-index.cfm?definition=myoglobin

History of Globins

- **Distribution of hemoglobin genes and pseudogenes** on the chromosomes reflect their evolution via duplication & divergence
- The expression of these genes follows a strict developmental pattern



The Genome of S. cerevisiae - Yeast

- One simplest **known** eukaryotic organisms
- Sequencing was completed in 1996
- Genome contains:
  - **12,057,500bp** of nuclear DNA
  - Distributed over **16 chromosomes**
    - Range in size **-** 1352 kbp chromosome IV to the 230 kbp chromosome I
  - 6172 predicted protein-coding genes
  - ~140 genes for rRNAs
  - 40 genes for small nuclear RNAs (snRNAs)
  - 275 tRNA genes
- **Denser in coding regions than known genomes of the more complex eukarya *Caenorhabditis elegans, Drosophila melanogaster* and human**
- Introns rare, and relatively small
  - Only **231 genes in yeast contain introns**
- There are fewer repeat sequences compared with more complex eukarya

Getting Data on Yeast

- Saccharomyces Genome Database
  - http://www.yeastgenome.org/
  - http://www.yeastgenome.org/download-data/sequence
  - http://browse.yeastgenome.org/fgb2/gbrowse/scgenome/
- Mips Saccharomyces cerevisiae genome database
  - http://mips.helmholtz-muenchen.de/genre/proj/yeast/
- Saccharoymces Genome Deletion Project
  - http://www-sequence.stanford.edu/group/yeast_deletion_project/deletions3.html
- Fungal Genomes – Wellcome Trust Sanger Institute
  - http://www.sanger.ac.uk/resources/downloads/fungi/

The Genome of C. elegans

- Completed 1998 - first full DNA sequence of a multicellular organism
- Contains 103,022,290 bp of DNA distributed on paired chromosomes:
  - I, II, III, IV, V and X
  - No Y chromosome
- Different genders in C. elegans
  - XX genotype, a self-fertilizing hermaphrodite
  - XO genotype, a male
- **~8 times larger yeast**
- 20,532 predicted genes ~**3x** the number of yeast
- Gene density relatively high for eukaryote
  - With ~1 gene/5 kb of DNA
  - Exons cover ~27% of the genome
  - The genes contain an average of five introns each
  - Approximately 25% of genes are in clusters of related genes

http://useast.ensembl.org/Caenorhabditis_elegans/Info/Annotation#assembly

C. elegans

- Many proteins - common other life forms
- Others are apparently specific to nematodes:
  - 42% of proteins have homologues outside the phylum;
  - 34% are homologous to proteins of other nematodes;
  - **24% have no known homologues outside C. elegans itself**

Distribution of *C. elegans* genes

Chromosome	Size (Mb)	Number of protein genes	Density of protein genes (kb/gene)	Number of tRNA genes
I	7.9	2803	5.06	13
II	8.5	3259	3.65	6
III	7.6	2508	5.40	9
IV	9.2	3094	5.17	7
V	9.8	4082	4.15	5
X	10.1	2631	6.54	3

FYI - WormBase

- Collaborative project to capture, curate and distribute information about *C.elegans* biology
- Began as ACeDB - database application software package developed by jointly Richard Durbin at the Sanger Institute and Jean-Thierry Mieq
- WormBase was originally started in 2000 as a way to make data in ACeDB easily accessible via a web-browser

http://www.wormbase.org/#012-3-6

The Genome of Homo Sapiens

- February 2001, **IHGSC & Celera Genomics** published separate drafts
- April 14, 2003 the finishing of genome was announced
  - within few days of 50th anniversary publication of Watson–Crick model of DNA
- ~3.2×10^9 bp
  - ~30x larger than genomes of *C. elegans* or *D. melanogaster*
  - One reason for disparity - coding sequences form **< 2% of human genome**; **repeat sequences over 50%**
- Small number of genes identified
  - ~20,700 suggests **alternative splicing patterns** make significant contribution to our **protein repertoire**
  - ~35% of genes have alternative splicing patterns

http://useast.ensembl.org/Homo_sapiens/Info/Annotation#assembly

Human Chromosomes

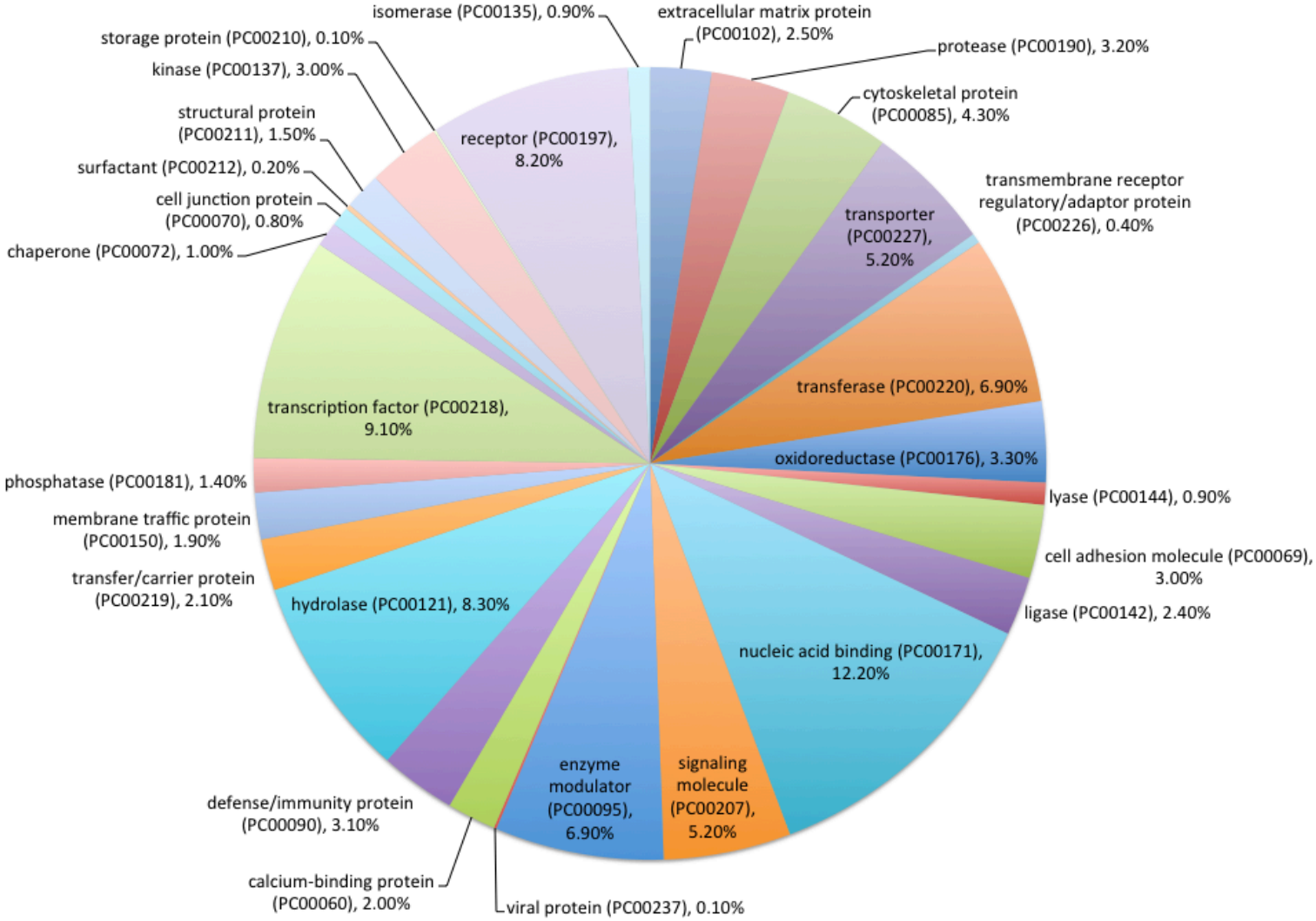
- Human genome is distributed over 22 chromosome pairs **plus** the X and Y chromosomes
- DNA contents of the autosomes range from 249 Mbp down to 48 Mbp
- The X chromosome contains 155 Mbp and the Y chromosome only 59 Mbp
- Good to know the relative sizes:
  - See here

    http://en.wikipedia.org/wiki/Human_genome

Human Genes

- Exons of human protein-coding genes relatively small compared w/ those in other known eukaryotic genomes
- **Introns are relatively long**
- As a result many protein-coding genes span long stretches of DNA
- For instance, the dystrophin gene, coding for a **3685** amino acid protein, is >2.4 Mbp long

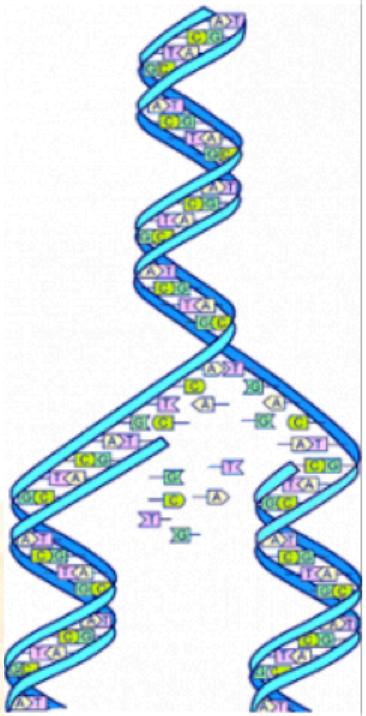
Human Protein Classes



http://www.pantherdb.org/chart/summary/pantherChart.jsp?filterLevel=1&chartType=1&listType=1&type=5&species=Homo%20sapiens

How Do We Know the HGP is Important?

Human Genome has a Facebook Page!!!



Info

Human Genome Project

Like

Interest

Description

From Wikipedia, the free encyclopedia

The **Human Genome Project** (**HGP**) is an international scientific research project with a primary goal of determining the sequence of chemical base pairs which make up DNA, and of identifying and mapping the approximately 20,000–25,000 genes of the human genome from both a physical and functional standpoint.

The project began in October 1990 and was initially headed by Ari Patrinos, head of the Office of Biological and Environmental Research in the U.S. Department of Energy's Office of Science. Francis Collins directed the National Institutes of Health National Human Genome Research Institute efforts. A working draft of the genome was announced in 2000 and a complete one in 2003, with further, more detailed analysis still being published. A parallel project was conducted outside of government by the Celera Corporation, which was formally launched in 1998. Most of the government-sponsored sequencing was performed in universities and research centers from the United States, the United Kingdom, Japan, France, Germany. The mapping of human genes is an important step in the development of medicines and other aspects of health care.

Source

22,488 like this

http://www.facebook.com/pages/Human-Genome-Project/108262042531075

Where Do I Begin to Get Human Genome Data?

Getting Data on Humans (1)

- NCBI – Human Genome Resources
  - http://www.ncbi.nlm.nih.gov/genome/guide/human/index.shtml
- Assembled Chromosomes
  - ftp://ftp.ncbi.nlm.nih.gov/genomes/H_sapiens/Assembled_chromosomes/
- International HapMap Project
  - http://snp.cshl.org/
- OMIM – Online Mendelian Inheritance in Man
  - http://www.ncbi.nlm.nih.gov/sites/entrez?db=omim
- NCBI medical genetics – GeneTests
  - http://www.ncbi.nlm.nih.gov/sites/GeneTests/?db=GeneTests

Getting Data on Humans (2)

- UCSC *Genome* Browser Home
  - http://genome.ucsc.edu/
- Gene Cards
  - http://www.genecards.org/
- HUGO – Gene Nomenclature Committee
  - http://www.genenames.org/
- 1000 Genomes
  - http://www.1000genomes.org/
- Exome Sequencing Project
  - http://evs.gs.washington.edu/EVS/

How Do We Capture Knowledge About Genomes?

Ontologies provide controlled, consistent vocabularies to describe concepts and relationships, thereby enabling

knowledge sharing" (Gruber 1993)

Capturing Knowledge

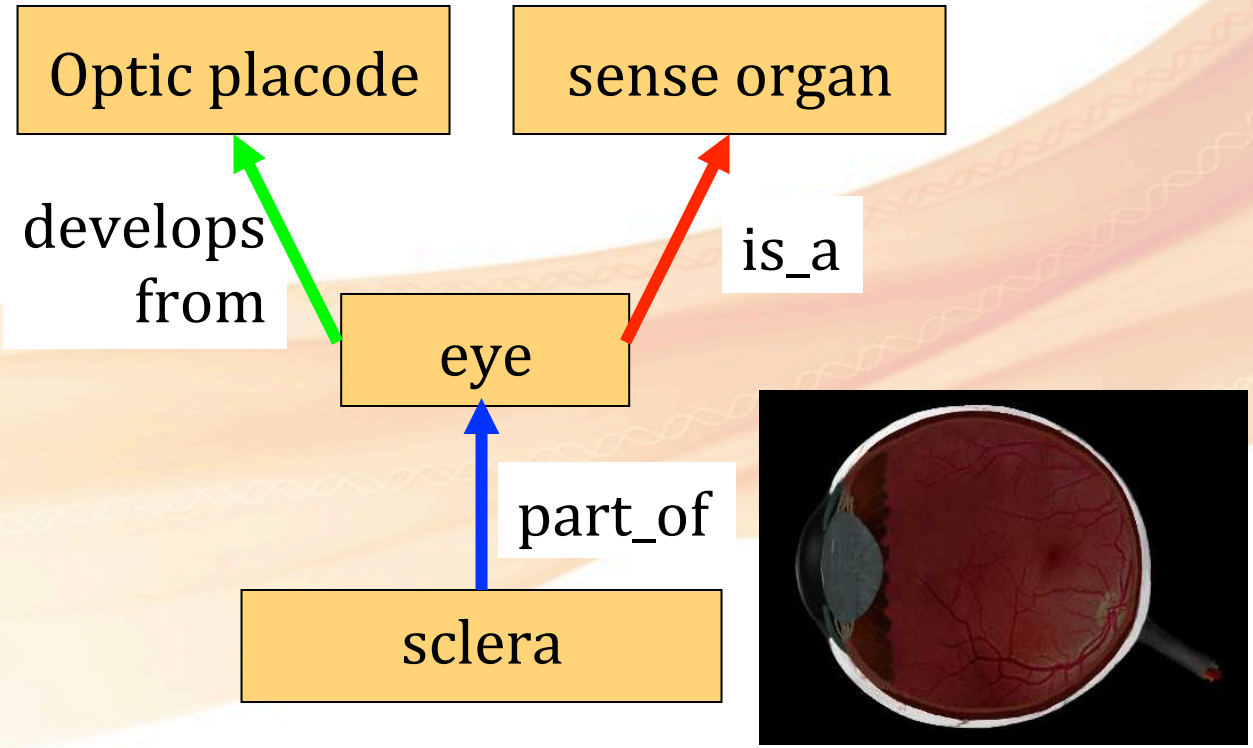
- If want capture knowledge for both humans and computer applications
- How would we start?
  - Need a set of vocabulary definitions
  - That capture a community's **knowledge of a domain**
- Use an **ontology**
  - 'An ontology may take a variety of forms:
    - It will include a ***vocabulary of terms***, and some *specification of their meaning*
    - Includes definitions and an indication of how concepts are inter-related
      - Which collectively impose a structure on the domain
      - And constrain the possible interpretations of terms'

A Biological Ontology is:

- A (machine) interpretable representation of some aspect of biological reality

what *kinds* of things exist?

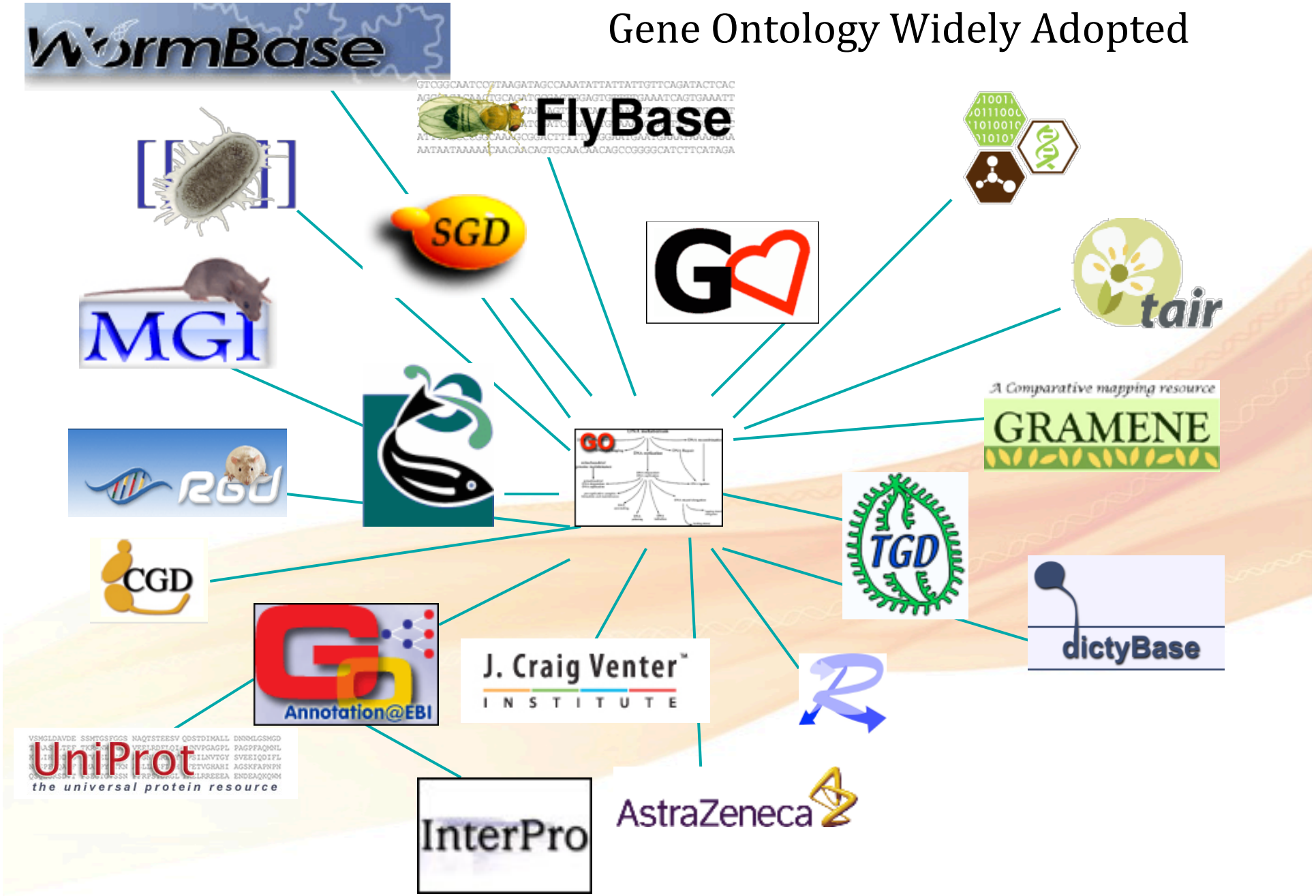
what are the *relationships* between these things?



What is Gene Ontology?

- The Gene Ontology (GO) project
  - http://www.geneontology.org
  - The GO is a database of terms for genes
- Major bioinformatics initiative
  - Aim of standardizing the representation of gene and gene product attributes across species and databases
  - Provides a controlled vocabulary of terms for describing gene product characteristics and gene product annotation data

Gene Ontology Widely Adopted



What are the Ontologies?

- Organized as three separate ontologies
  - is located within a certain **Cellular components**
  - has a particular **Molecular functions**
  - is involved in a particular **Biological processes**

GO Term

- **As determined by a particular method**

Evidence Code

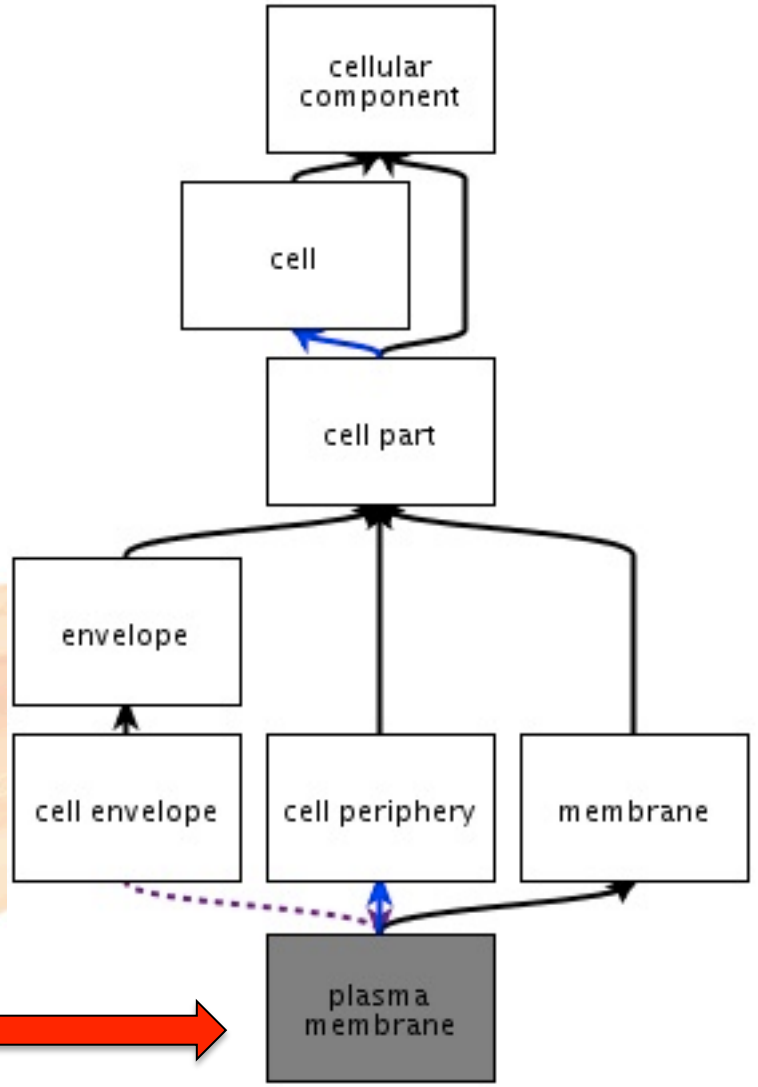
- **As described in a particular reference**

Reference Code

- Each gene or gene product may
  - Have more than one molecular function
  - Take part in more than one biological process
  - Act in more than one cellular component

GO Annotations

- Controlled vocabulary
  - Everyone uses the same terms
  - Terms have 7 digit IDs that computers can understand
- Relationships between terms



GO:0005886

Anatomy of an Annotation

- Object (previously mentioned)
- **GO Term** from most recent GO
  - GO Term Qualifier (optional)
    - NOT, colocalizes_with, contributes_to
- **Evidence Code**: IDA, IPI, IEP, IGI, ISS, IEA, TAS, NAS, or IC
  - Evidence Code Qualifier (required for some codes)
    - Used in combination with IPI, IMP, IGI, and ISS
      - Seq_ID or DB_ID required
- **Reference Code**: literature or database specific reference
  - DB_ID or PMID

Adopted from Harold Drabkin

Anatomy of an Annotation - NOT

- **Used** with terms from any of 3 ontologies
- Used to make an explicit note - gene product is not associated with the GO term
- Important in cases where associating a GO term with a gene product should be avoided
- For example:
  - If a protein has sequence similarity to an enzyme (whose activity is GO:nnnnnnn)
  - But has been shown experimentally not to have the enzymatic activity, it can be annotated as **NOT** GO:nnnnnnn
- **Used** when a GO term might otherwise be expected to apply to a gene product, but an experiment, sequence analysis, etc. proves otherwise

Anatomy of an Annotation - colocalizes_with

- **Used** only with cellular component terms
- Gene products - transiently or peripherally associated with an organelle or complex may be annotated to the relevant cellular component term using **colocalizes_with**
- Also used in cases where the resolution of an assay is not accurate enough to say that **the gene product** is a bona fide component member

- Example (from *Schizosaccharomyces pombe*):
  - Clp1p relocalizes from the nucleolus to the spindle and site of cell division; i.e. it is associated transiently with the spindle pole body and the contractile ring (evidence from GFP fusion)
  - Clp1p is annotated to spindle pole body - GO:0005816 and contractile ring - GO:0005826, using the c**olocalizes_with** qualifier in both cases

Anatomy of an Annotation – contribues_to

- **Used** only with molecular function terms
- As noted, individual gene product that is part of a complex can be annotated to terms that describe the function of the complex
- Many such function annotations should use the qualifier contributes_to:
- Especially useful for molecular function annotations in cases where a **complex** has an activity, but not all of the individual subunits do
- Example
  - there may be a known catalytic subunit and one or more additional subunits
  - or the activity may only be present when the complex is assembled
- Molecular function annotations of complex subunits that are not known to possess the activity of the complex must include the entry **contributes_to** in the Qualifier column

What are the Ontologies?

- Organized as three separate ontologies
  - **Cellular components**
  - **Molecular functions**
  - **Biological processes**
- Each gene or gene product may
  - Have more than one molecular function
  - Take part in more than one biological process
  - Act in more than one cellular component

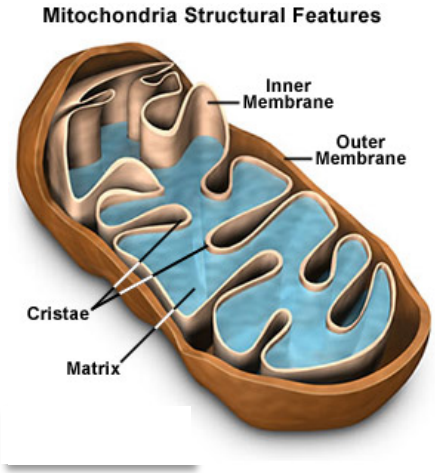
Three Separate Ontologies

Three Separate Ontologies

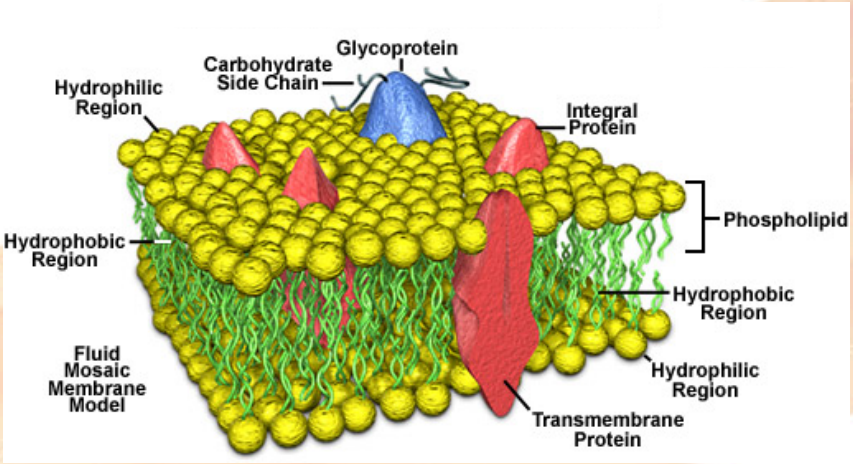
- Cellular component ontology = location or complex
  - Where does gene product act?
  - Subcellular structures, locations, and macromolecular complexes
    - examples: *nucleus*, *telomere*, and *RNA polymerase II holoenzyme*
- Molecular function ontology = elemental activity/task
  - Which molecular function does gene product have?
  - Tasks performed by individual gene products
    - examples: *carbohydrate binding* and *ATPase activity*
- Biological process ontology = biological goal or objective
  - Which process is a gene product involved in?
  - Broad biological goals that accomplished by ordered assemblies of molecular functions
    - examples: *mitosis* or *purine metabolism*

Cellular Component

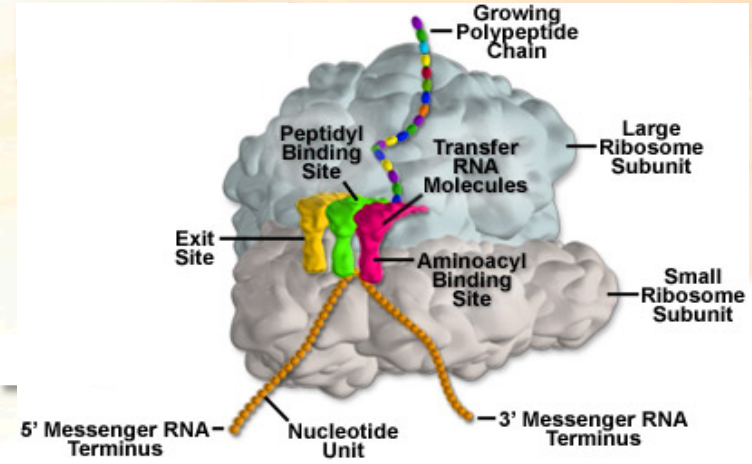
Where does gene product act?



GO:0005739 mitochondrion



GO:0009274 peptidoglycan-based cell wall

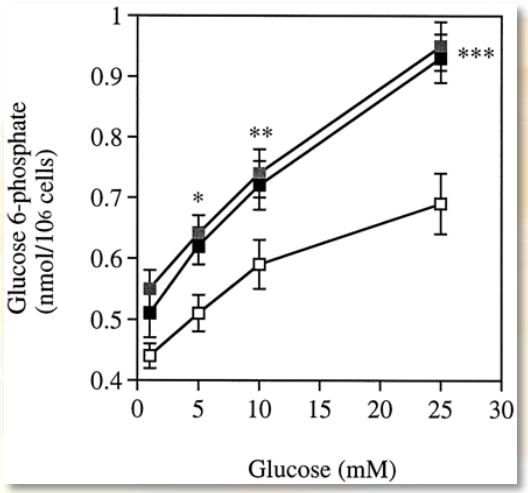


GO:0005840 ribosome

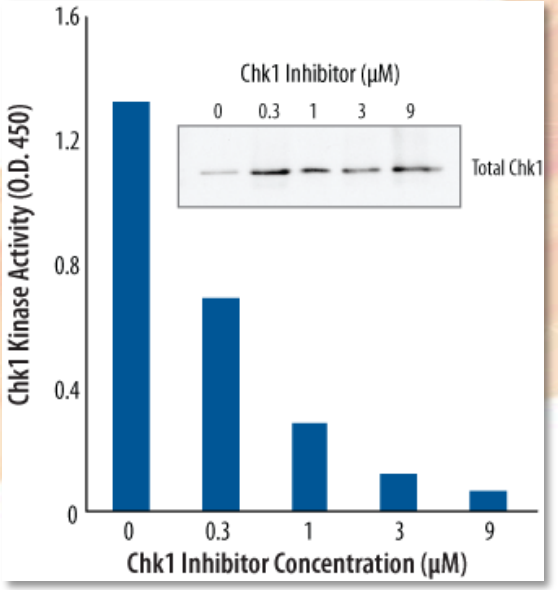
Molecular Function

Which molecular function does gene product have?

GO:0004347 glucose-6-phosphate isomerase activity



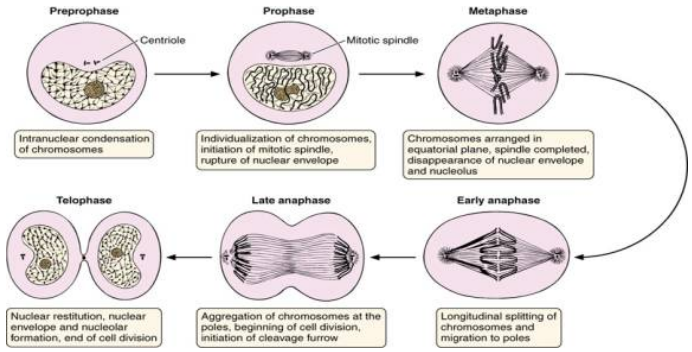
GO:0004396 hexokinase activity



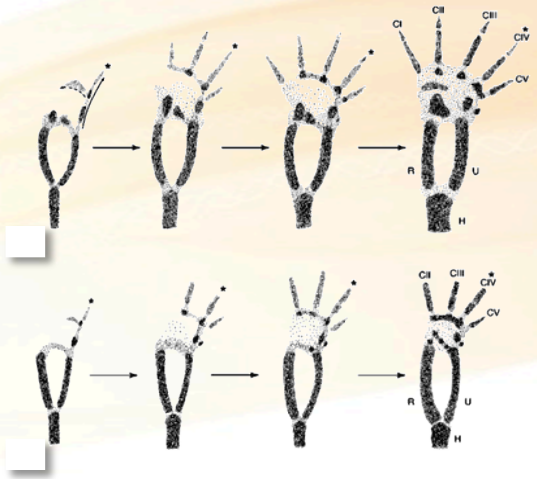
GO:0016301 Kinase activity

Biological Process

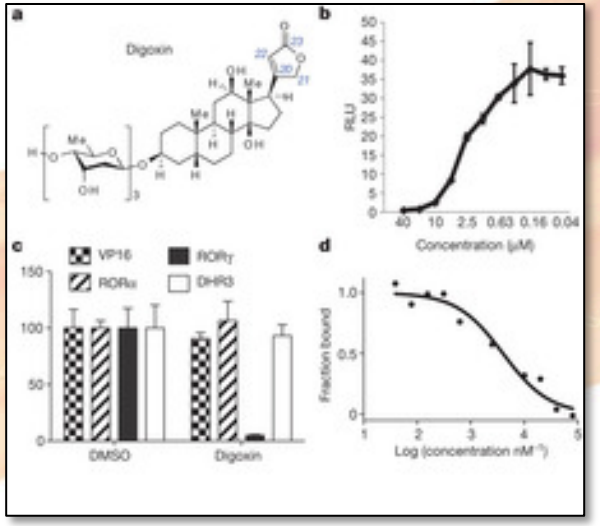
which process is a gene product is involved in



GO:0051301 cell division



GO:0060173 limb development



GO:0006351 transcription, DNA dependent

What Does an Ontology Do?

- Captures knowledge
- Creates a shared understanding – between humans and for computers
- Makes knowledge **machine processable**
- Makes meaning explicit – by definition and context
- Bases annotations on Evidence

The Evidence of a GO Annotation

- GO annotation consists of a **GO term** associated with a **specific reference**
- Reference describes the analysis upon which the association between a specific **GO term** and **gene product** is based
- Each annotation must also include:
  - **Evidence code -** how the annotation to a particular term is supported
  - Although evidence codes reflect the type of analysis described in the reference
    - Which **supports the GO term to gene product association**
    - They are **not necessarily** a **classification of types of experiments/ analyses**

http://www.geneontology.org/GO.evidence.shtml

The Evidence – Four General Categories

- **Annotations based on Evidence**
- Only Inferred from Electronic Annotation (**IEA**) = **not** assigned by a curator
- **Manually-assigned evidence codes** fall into four general categories:
  - **Experimental**
    - **cited paper displayed results from a physical characterization of a gene or gene product – supports association of a GO term**
  - **Computational analysis**
    - ***in silico* analysis of the gene sequence and/or other data as described in the cited reference**
  - Author statements
    - statement made by the author(s) in the reference cited
  - Curatorial statements
    - curatorial judgement that does not fit into one of the other evidence code classifications

http://www.geneontology.org/GO.evidence.shtml

GO Evidence Codes - Experimental

Code	Definition	Type
EXP	Inferred from Experiment	Experimental (PARENT)
IDA	Inferred from Direct Assay	Experimental
IPI	Inferred from Physical Interaction	Experimental
IMP	Inferred from Mutant Phenotype	Experimental
IGI	Inferred from Genetic Interaction	Experimental
IEP	Inferred from Expression Pattern	Experimental

- IDA:
  - Enzyme assays
  - In vitro reconstitution
  - Immunofluorescence
  - Cell refraction

http://www.geneontology.org/GO.evidence.shtml

GO Evidence Codes - Computational

Code	Definition	Type
ISS	**Inferred from Sequence or structural Similarity**	**Computational**
ISO	Inferred from Sequence Ortholog	Computational
ISA	Inferred from Sequence	Computational
ISM	Inferred from Sequence Model	Computational
IGC	**Inferred from Genomic Context**	**Computational**
IBA	Inferred from Biological aspect of Ancestor	Computational
IBD	Inferred from Biological aspect of Descendant	Computational
IKR	Inferred from Key Residues	Computational
IRD	Inferred from Rapid Divergence	Computational
RCA	inferred from Reviewed Computational Analysis	Computational

http://www.geneontology.org/GO.evidence.shtml

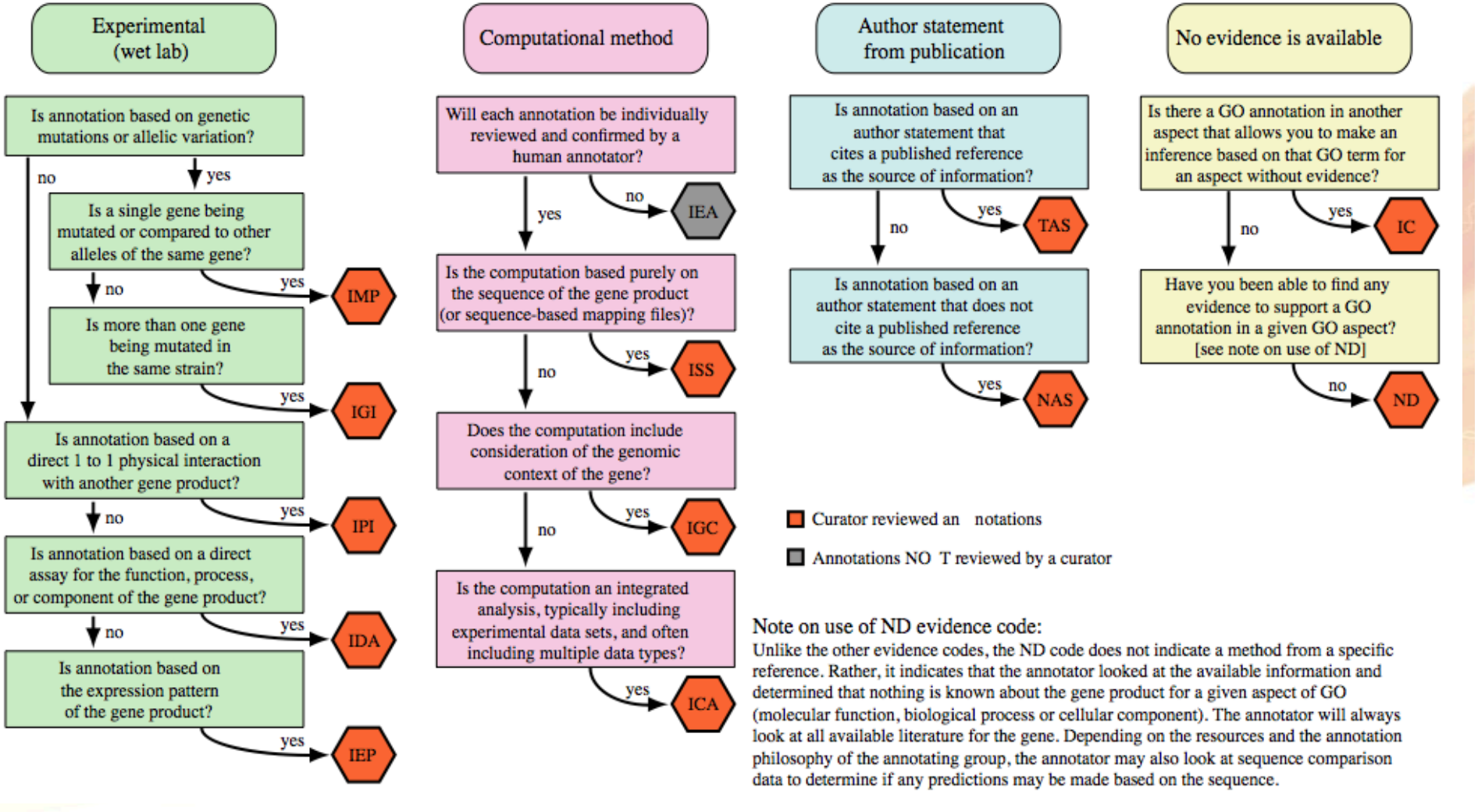
GO Evidence Codes – Author/Curatorial

Code	Definition	Type
TAS	Traceable Author Statement	Author
NAS	Non-traceable Author Statement	Author
IC	Inferred by Curator	Computational
ND	No biological Data available evidence code	Computational

- TAS:
  - In the literature source the original experiments are referenced.

http://www.geneontology.org/GO.evidence.shtml

What type of Evidence is the Annotation Based On – Decision Treee



http://www.geneontology.org/GO.evidence.tree.shtml

Annotation Strategies

- Electronic (IEA)
  - Good first pass
  - Usually based on some sort of sequence comparison – Mapping of external concept to GO terms
    - InterPro2GO
    - SPKW2GO
    - HAMAP2GO
    - EC2GO
    - SPSL2GO
    - All these annotations have IEA evidence code
- Manual (curator reviewed, mostly inferred form literature)

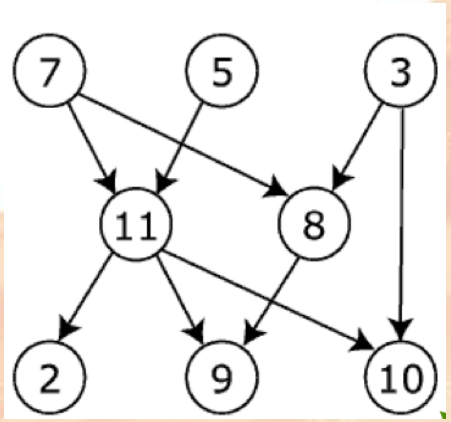
Literature Selection – Mouse Example

- A paper is selected for GO curation of a **mouse** gene product if:
  - A paper provides **direct** experimental evidence for the **normal** function, process, or cellular location of a:
    - **mouse** gene product (IDA, IMP, IGI, IPI, IEP)
    - non-mouse gene product **AND** the paper present homology data to a mouse gene product (ISS)

Structure of Ontologies

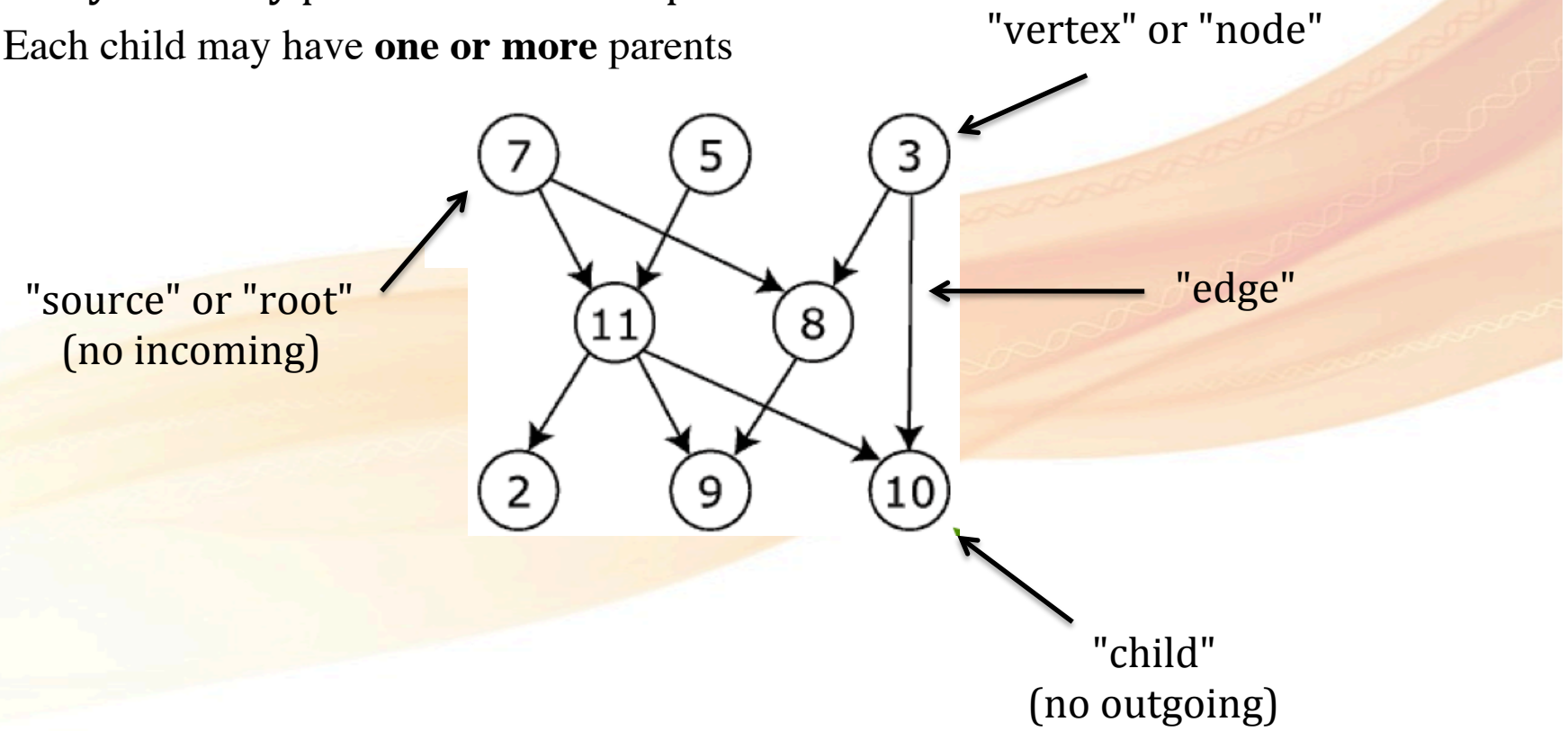
Structure of Ontologies

- Shows relation between different terms - one term may be more specific description of another more general term
- Represented as **Directed Acyclic Graph (DAG)**
  - Similar to hierarchies
  - Allows a child node to have more than one parent
  - Levels represent specifity of the terms
- Every child-term is a member of its parent-term



Directed Acyclic Graph (DAG)

- No cycles
- In between a tree and a graph
- **Many-to-many** parental relationship
- Each child may have **one or more** parents



True Path Rule

- The path from a child term all the way up to its top-level parent(s) must always be true

cell

Ⓟ cytoplasm

Ⓟ chromosome

ⓘ nuclear chromosome

ⓘ cytoplasmic chromosome

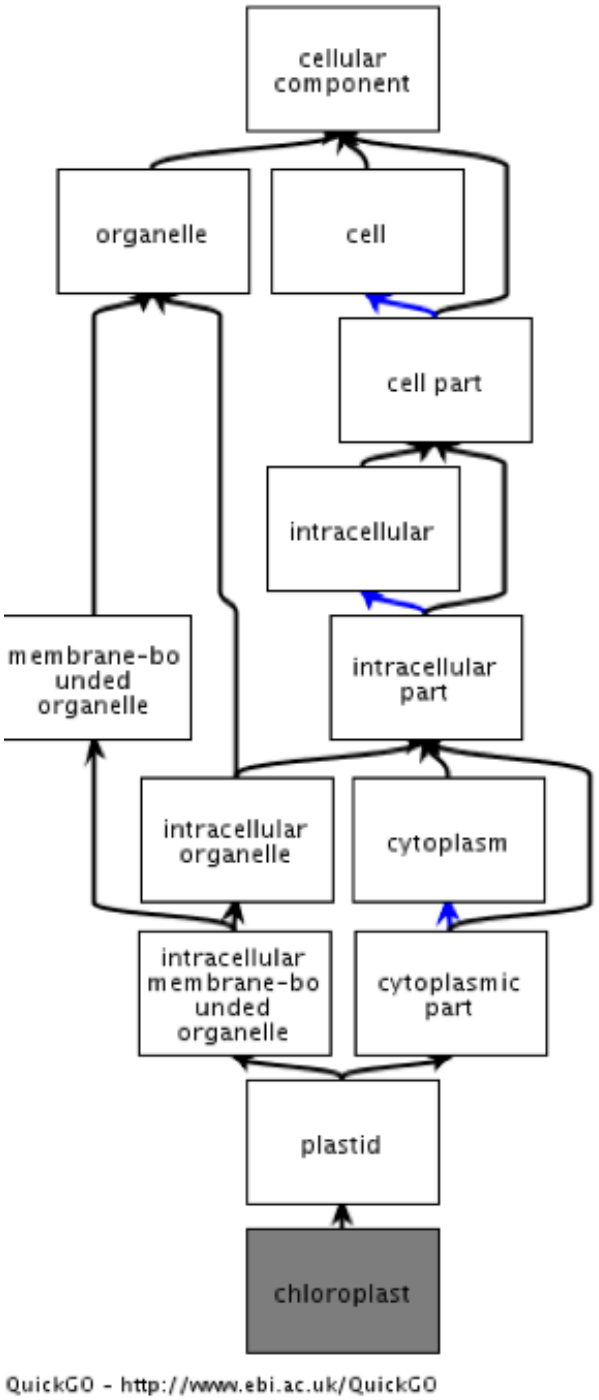
ⓘ mitochondrial chromosome

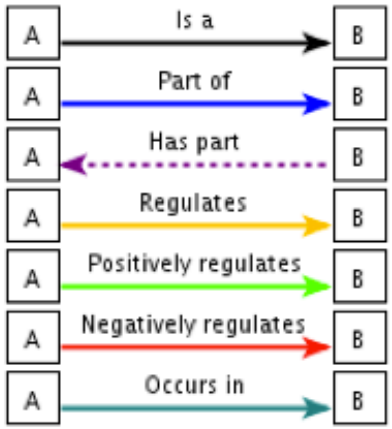
Ⓟ nucleus

Ⓟ nuclear chromosome

is-a ⓘ

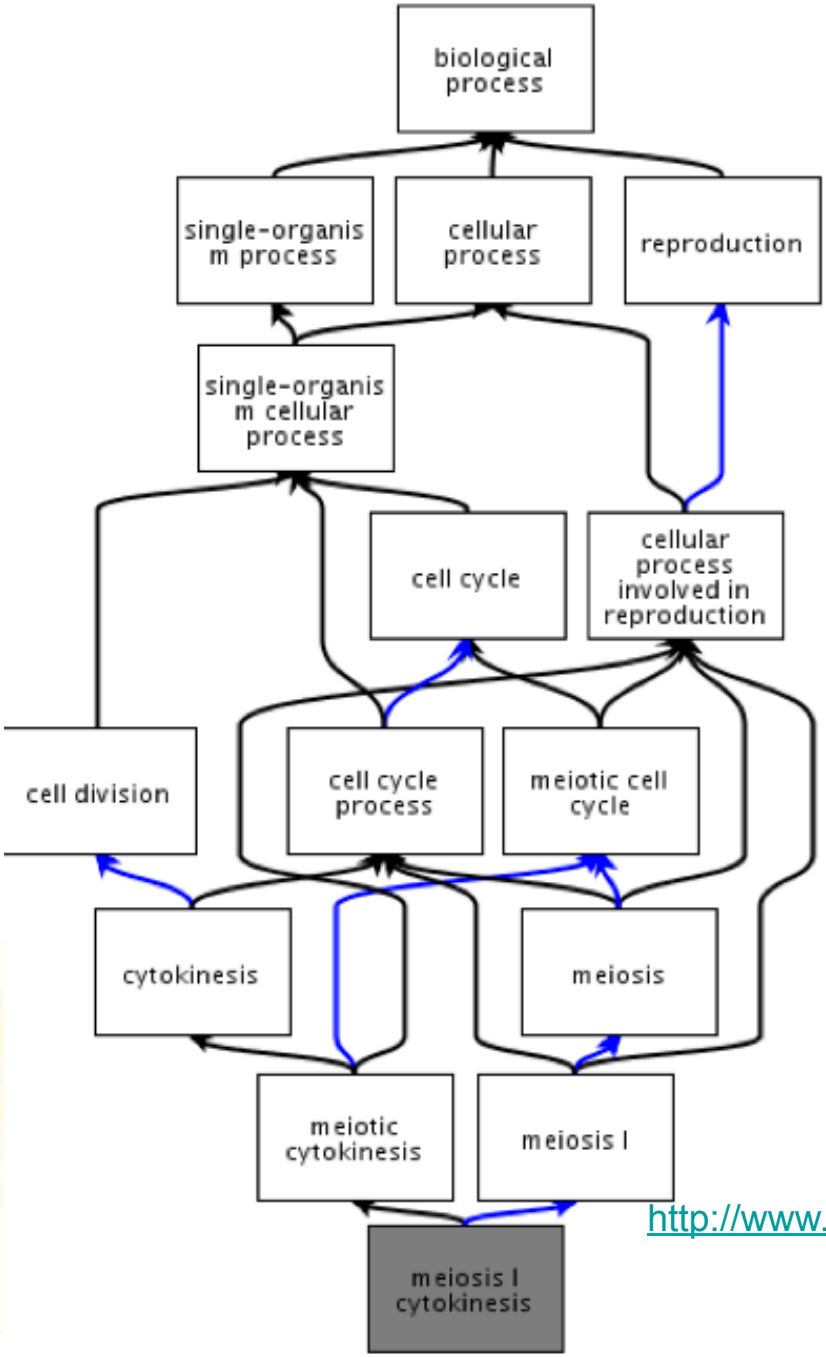
part-of Ⓟ

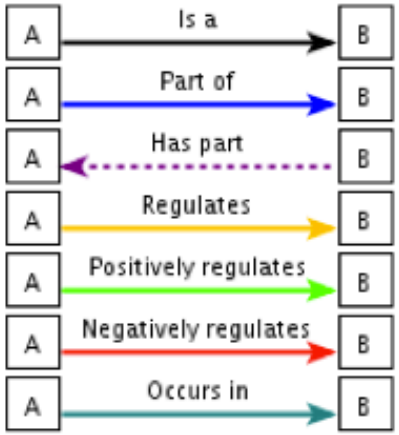




GO:0009507 chloroplast

http://www.ebi.ac.uk/QuickGO/GTerm?id=GO:0009507#term=ancchart





GO:0007110 meiosis I cytokinesis

http://www.ebi.ac.uk/QuickGO/GTerm?id=GO:0007110#term=ancchart

Anatomy of a GO term

id: GO:0006094	**unique GO ID**
name: gluconeogenesis	term name
namespace: process	ontology
def: The formation of glucose from noncarbohydrate precursors, such as pyruvate, amino acids and glycerol. [http://cancerweb.ncl.ac.uk/omd/index.html]	definition
exact_synonym: glucose biosynthesis	synonym
xref_analog: MetaCyc:GLUCONEO-PWY	database ref
is_a: GO:0006006 is_a: GO:0006092	parentage

Here's the first entry (of the ~35K) in the GO text version (with all three parts intermixed):

```
[Term]
id: GO:0000001
name: mitochondrion inheritance
namespace: biological_process
def: "The distribution of mitochondria, including
the mitochondrial genome, into daughter cells
after mitosis or meiosis, mediated by
interactions between mitochondria and the
cytoskeleton." [GOC:mcc, PMID:10873824, PMID:
11389764]
synonym: "mitochondrial inheritance" EXACT []
is_a: GO:0048308 ! organelle inheritance
is_a: GO:0048311 ! mitochondrion distribution
```

You can also get the GO as RDF XML, or as a MySQL database.

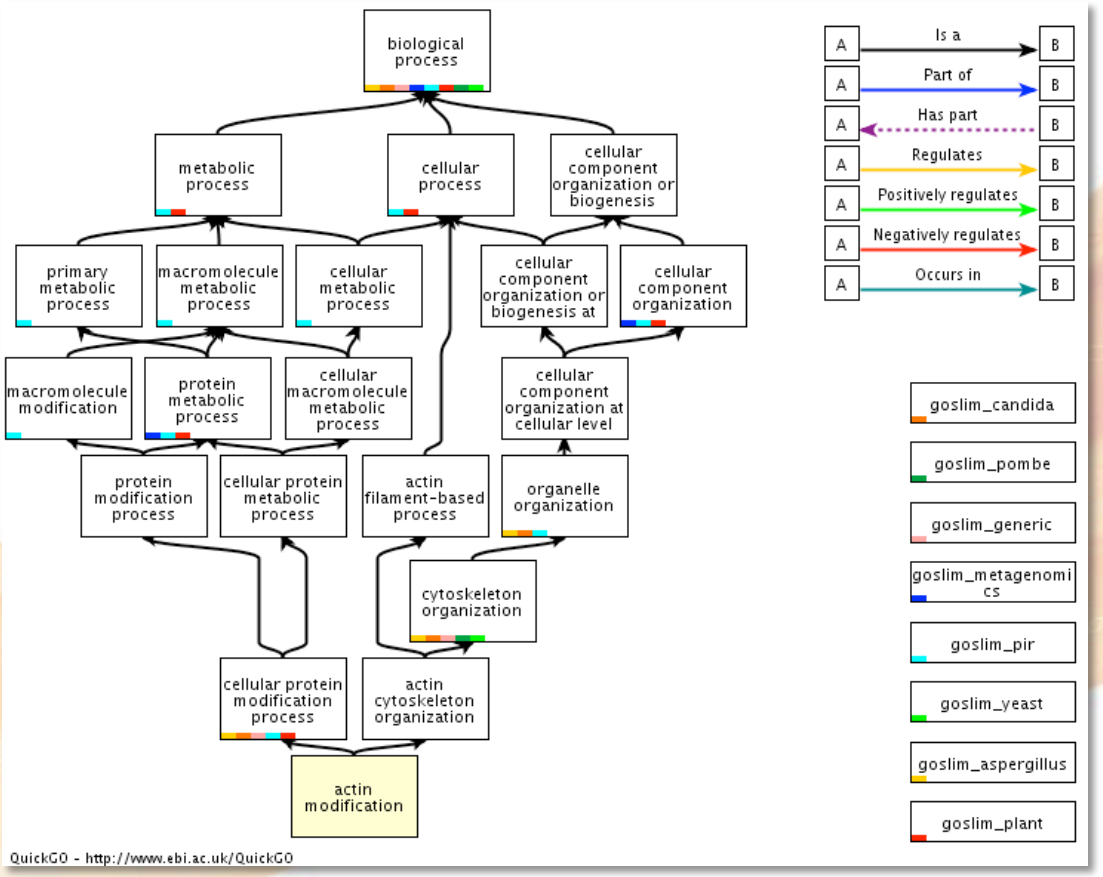
http://www.geneontology.org/GO.downloads.ontology.shtml

Example Search: GO:0030047

actin modification

- Accession: **GO:0030047**
- Ontology: **Biological Process**
- Synonyms: alt_id: **GO:0007013**
- Definition: **Covalent modification of an actin molecule.**
  Source: GOC:mah
- Comment: **None**
- Subset: **None**
- Community **Add usage comments for this term on the GONUTS wiki**

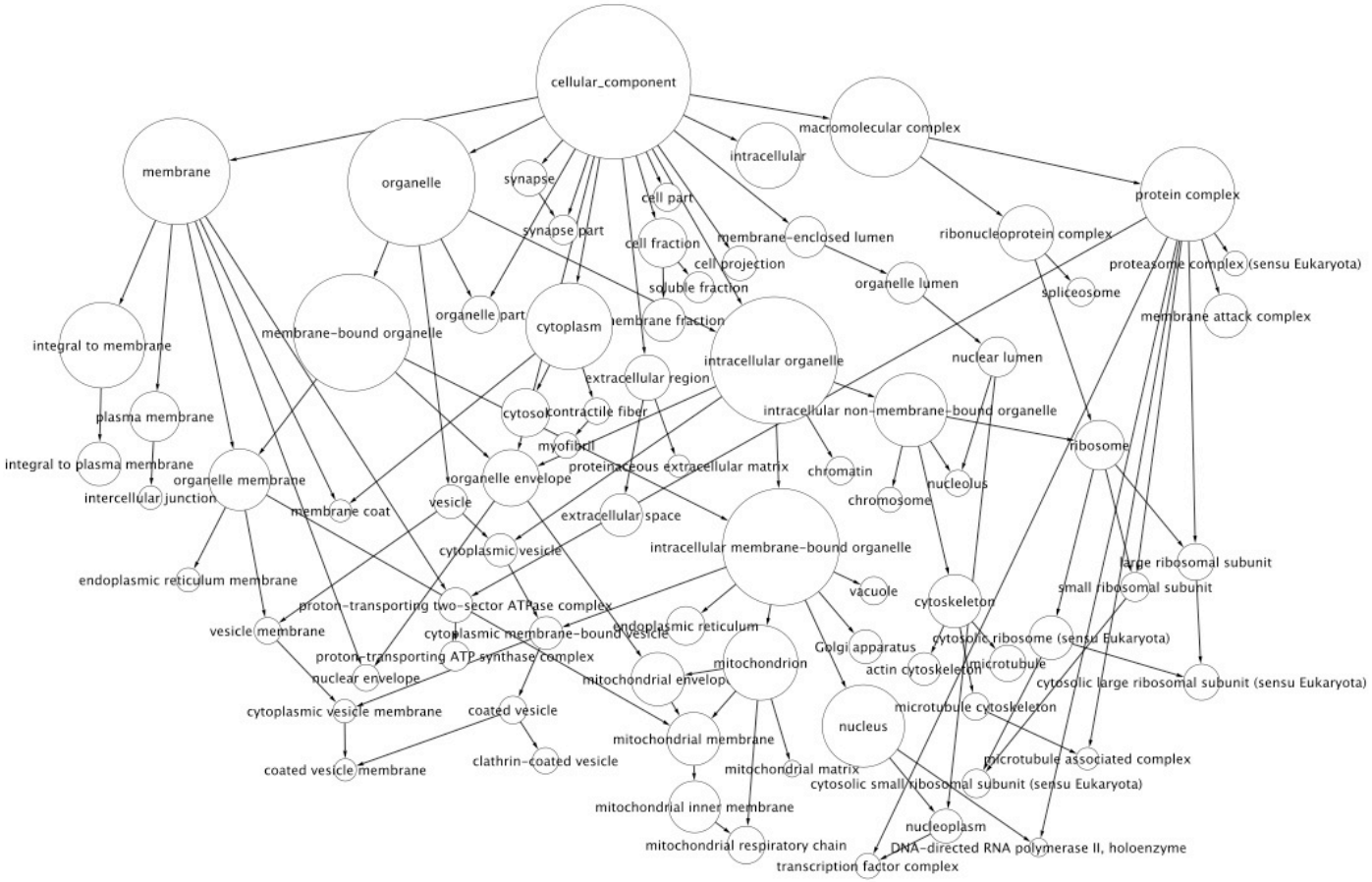
- **Annotation to GO:0020047 automatically annotates to all of its parents, thus product is annotated both:**
  - **protein modification**
  - **actin cytoskeleton organization**



Graph View

http://www.ebi.ac.uk/QuickGO/

More Complex Example



Directed Acyclic Graph of Gene Ontology Terms applied to the A. burtoni EST set on the second generation micorarray.

Who Uses GO?

NIH funded Experimental Research that Uses GO

- National Institute on Alcohol Abuse and Alcoholism (NIAAA)
- National Institute on Aging (NIA)
- National Institute of Allergy and Infectious Diseases (NIAID)
- National Institute of Arthritis and Musculoskeletal and Skin Diseases (NIAMS)
- National Center for Complementary and Alternative Medicine (NCCAM)
- National Cancer Institute (NCI)
- National Institute on Drug Abuse(NIDA)
- National Institute on Deafness and Other Communication Disorders (NIDCD)
- National Institute of Dental &Craniofacial Research (NIDCR)
- National Institute of Diabetes andDigestive and Kidney Diseases(NIDDK)
- National Institute of BiomedicalImaging and Bioengineering (NIBIB)
- National Institute of Environmental Health Sciences (NIEHS)

many more…

Companies

- NLP & Ontology Products
  - BioWisdom
  - IBM
- Array Products and data analysis
  - Affymetrix
  - Spotfire
- Products that use GO
  - Clontech
  - Inpharmatica
- Reagents& services
  - Sigma
  - Abnova
- Big Pharmas
  - AstraZeneca

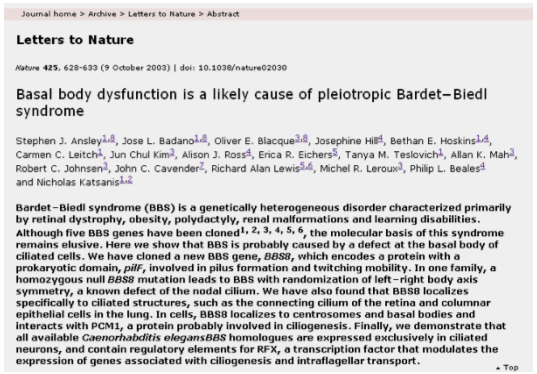
Let's Take 5 Minutes…

The Reference Genome Annotation Project

We are in the middle of an explosion of genomic information

http://www.geneontology.org/GO.refgenome.shtml

Who Classically Makes Functional Annotations?



Journal home > Archive > Letters to Nature > Abstract

Letters to Nature

Nature **425**, 628-633 (9 October 2003) | doi: 10.1038/nature02030

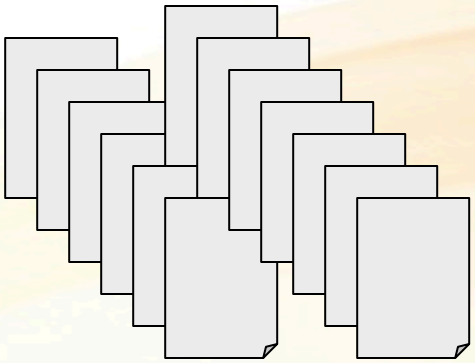
Basal body dysfunction is a likely cause of pleiotropic Bardet–Biedl syndrome

Stephen J. Ansley[1,8], Jose L. Badano[1,8], Oliver E. Blacque[3,8], Josephine Hill[4], Bethan E. Hoskins[1,5], Carmen C. Leitch[1], Jun Chul Kim[3], Alison J. Ross[3], Erica R. Eichers[3], Tanya M. Teslovich[1], Allan K. Mah[3], Robert C. Johnsen[3], John C. Cavender[2], Richard Alan Lewis[3,6], Michel R. Leroux[3], Philip L. Beales[4] and Nicholas Katsanis[1,2]

Bardet–Biedl syndrome (BBS) is a genetically heterogeneous disorder characterized primarily by retinal dystrophy, obesity, polydactyly, renal malformations and learning disabilities. Although five BBS genes have been cloned[1, 2, 3, 4, 5, 6], the molecular basis of this syndrome remains elusive. Here we show that BBS is probably caused by a defect at the basal body of ciliated cells. We have cloned a new BBS gene, *BBS8*, which encodes a protein with a prokaryotic domain, *pilF*, involved in pilus formation and twitching mobility. In one family, a homozygous null *BBS8* mutation leads to BBS with randomization of left–right body axis symmetry, a known defect of the nodal cilium. We have also found that BBS8 localizes specifically to ciliated structures, such as the connecting cilium of the retina and columnar epithelial cells in the lung. In cells, BBS8 localizes to centrosomes and basal bodies and interacts with PCM1, a protein probably involved in ciliogenesis. Finally, we demonstrate that all available *Caenorhabditis elegans BBS* homologues are expressed exclusively in ciliated neurons, and contain regulatory elements for RFX, a transcription factor that modulates the expression of genes associated with ciliogenesis and intraflagellar transport.

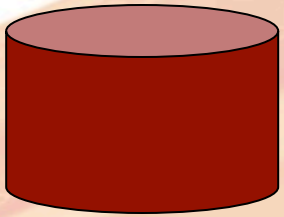
▲ Top

Literature



Datasets

Biocurators
(rate limiting)



Database

With More and More Genomes Being Sequenced

- GO has Limited resources:
  - To manually annotate the growing number of sequenced genomes
  - **Automatic annotation** will be the method of choice for many groups
- GO Consortium coordinated an effort to maximize and optimize the GO annotation of a large and representative set of key genomes
  - 'Reference Genomes'
  - Goal:
    - Completely annotate 12 reference genomes
    - These annotations = effectively **seed** the **automatic** annotation efforts of **other genomes**

Reference Species and Databases

- *Arabidopsis thaliana* (The *Arabidopsis* Information Resource (TAIR))
- *Caenorhabditis elegans* (WormBase)
- *Danio rerio* (zebrafish; Zebrafish Information Network (ZFIN))
- *Dictyostelium discoideum* (dictyBase)
- *Drosophila melanogaster* (FlyBase)
- *Escherichia coli* (PortEco)
- *Gallus gallus* (AgBase)
- *Homo sapiens* (
  human UniProtKB-Gene Ontology Annotation [UniProtKB-GOA] @ EBI)
- *Mus musculus* (Mouse Genome Informatics)
- *Rattus norvegicus* (Rat Genome Database (RGD))
- *Saccharomyces cerevisiae* (*Saccharomyces* Genome Database (SGD))
- *Schizosaccharomyces pombe* (Pombase)

Overview of Project Goal and Strategy

- Goals of the Reference Genome Project are:
  - Provide:
    - Set of comprehensive experimental GO annotations for all gene products in all 12 Reference Genomes
    - Tools for **using these annotations** to infer GO annotations for all fully sequenced genomes
- Evolutionary relationships are the "glue" in the Reference Genome Project
- Related genes across all twelve Reference Genomes are curated simultaneously, providing:
  - Better annotations
  - More annotations
  - Improvements to GO

Better Annotations

- Each model organism has unique strengths for probing gene function
- Bringing this information together helps to interpret experimental results, which improves the accuracy and consistency of annotations

More Annotations

- Homology relationships allow accurate inference of functions for genes that have not been characterized experimentally

Improvements in the Gene Ontology

- Cross-organism discussion about annotations frequently leads to new terms being added to the Gene Ontology

FYI - Curation Process

- 1. Identify the initial set of target genes in (typically) one species
  - Genes are selected that belong to either one of the four following categories:
    - Orthologs of human disease genes
    - Topical or 'hot' genes
    - Genes conserved from E. coli to human but currently lacking GO annotation
    - Genes involved in biochemical and/or signaling pathways

FYI - Curation Process

- 2. Identify the ortholog(s)/homolog(s) of the selected target genes in all Reference Genome species, from phylogenetic trees in the PANTHER database
  - Not all species may have orthologs/homologs to selected genes
- 3. Curators from each model organism database collect available literature about the genes in their respective organism
- 4. Curators assign GO terms based on experimental data
- 5. Review existing GO annotations to make sure they conform to agreed GO annotation standards
- 6. Overlay all annotations on the phylogenetic tree of the gene family. Annotations are reviewed for consistency, and modified if necessary. Ancestral nodes in the tree are annotated, allowing reliable, traceable inference of annotations based on homology. These processes are carried out using the PAINT (Phylogenetic Annotation INference Tool) software operating on the trees in the PANTHER database.

FYI - Curation Process

- 6. Overlay all annotations on the phylogenetic tree of the gene family
  - Annotations are reviewed for consistency, and modified if necessary
  - Ancestral nodes in the tree are annotated, allowing reliable, traceable inference of annotations based on homology
  - These processes are carried out using the PAINT (Phylogenetic Annotation INference Tool) software operating on the trees in the PANTHER database

How Does this Project Differ from Standard GO Annotation?

- The main results of this process are:
  - Additional quality assurance of experimental GO annotations by viewing each annotation in the context of annotations for related genes
  - A set of high-quality inferred GO annotations derived from the annotated phylogenetic trees
  - A fully traceable evidence trail for all annotations, both experimental and inferred
- The reference genome databases have agreed to follow more stringent guidelines than those used for standard GO annotation

How Can GO be Used?

1- Finding Enriched GO terms

- A comparison between two gene sets
  - Looking for the GO terms that are enriched in one of the gene sets and relatively depleted in the other
- For every GO term, a p-value can be calculated from the following table

	# of genes associated with this GO term	# of genes not associated
Gene List 1	Obs_{11}	Obs_{12}
Gene List 2	Obs_{21}	Obs_{22}

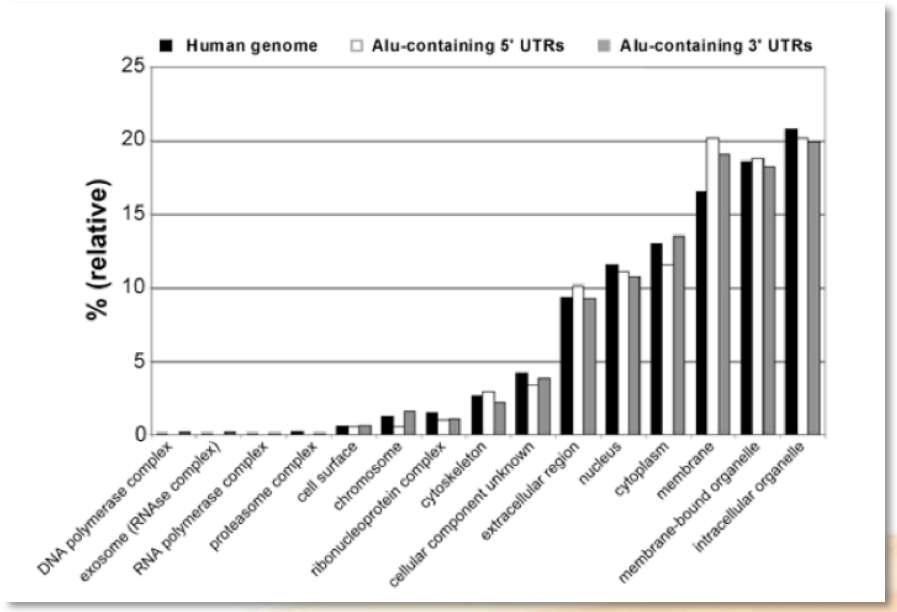
The null hypothesis:

$$\frac{Obs_{11}}{Obs_{21}} = \frac{Obs_{12}}{Obs_{22}}$$

The alternative:

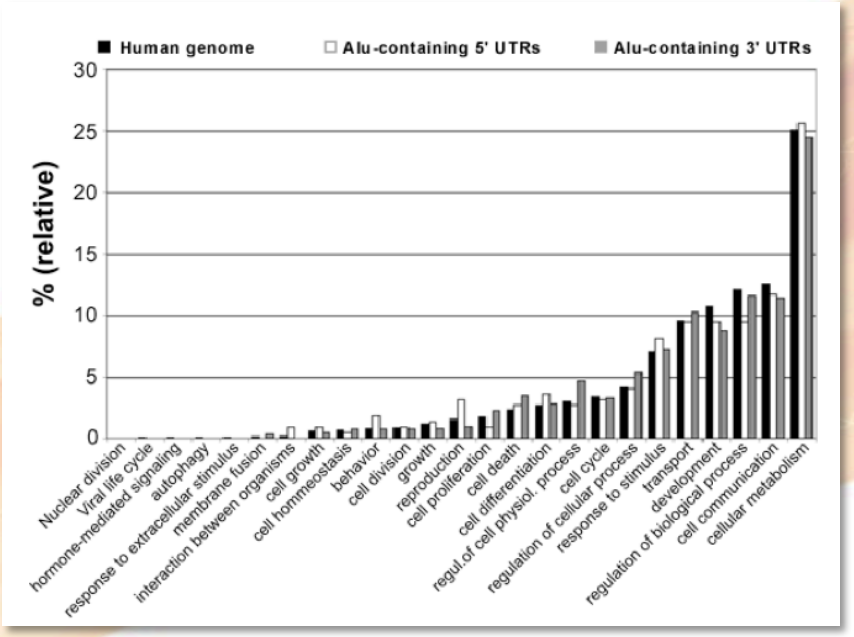
$$\frac{Obs_{11}}{Obs_{21}} \neq \frac{Obs_{12}}{Obs_{22}}$$

2- Analysis of mRNA



Biological process

Cellular Component



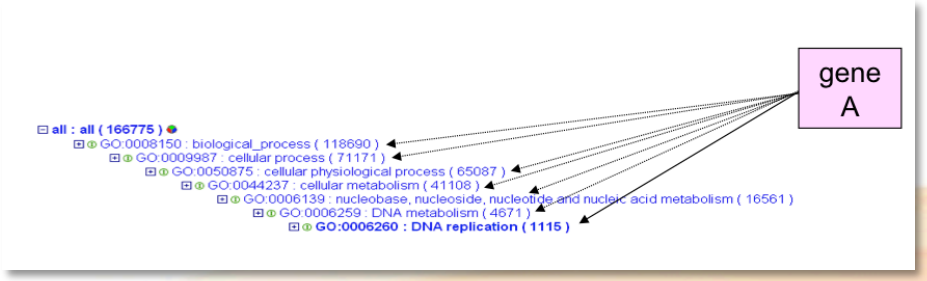
What's this show?

The results show there is no significant difference between the human genome and the two mRNA assemblies in terms of distribution into the different GO categories

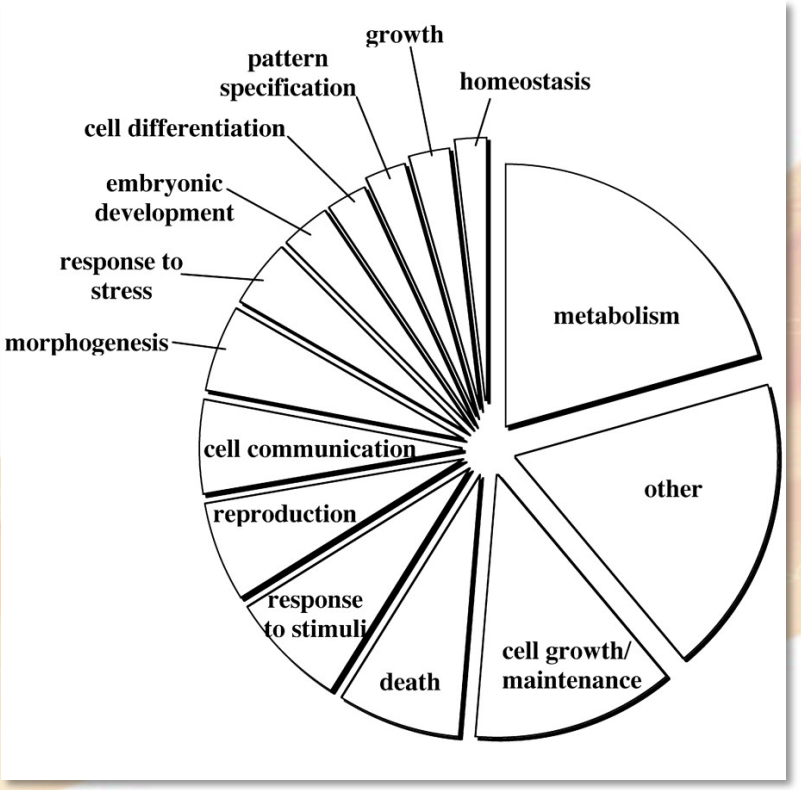
http://www.unige.ch/sciences/biologie/bicel/Strub/researchAlu.html

3 - Broad Overview of Gene Sets

- GO isn't just a flat list of biological terms
  - Remember, terms are related within the ontology

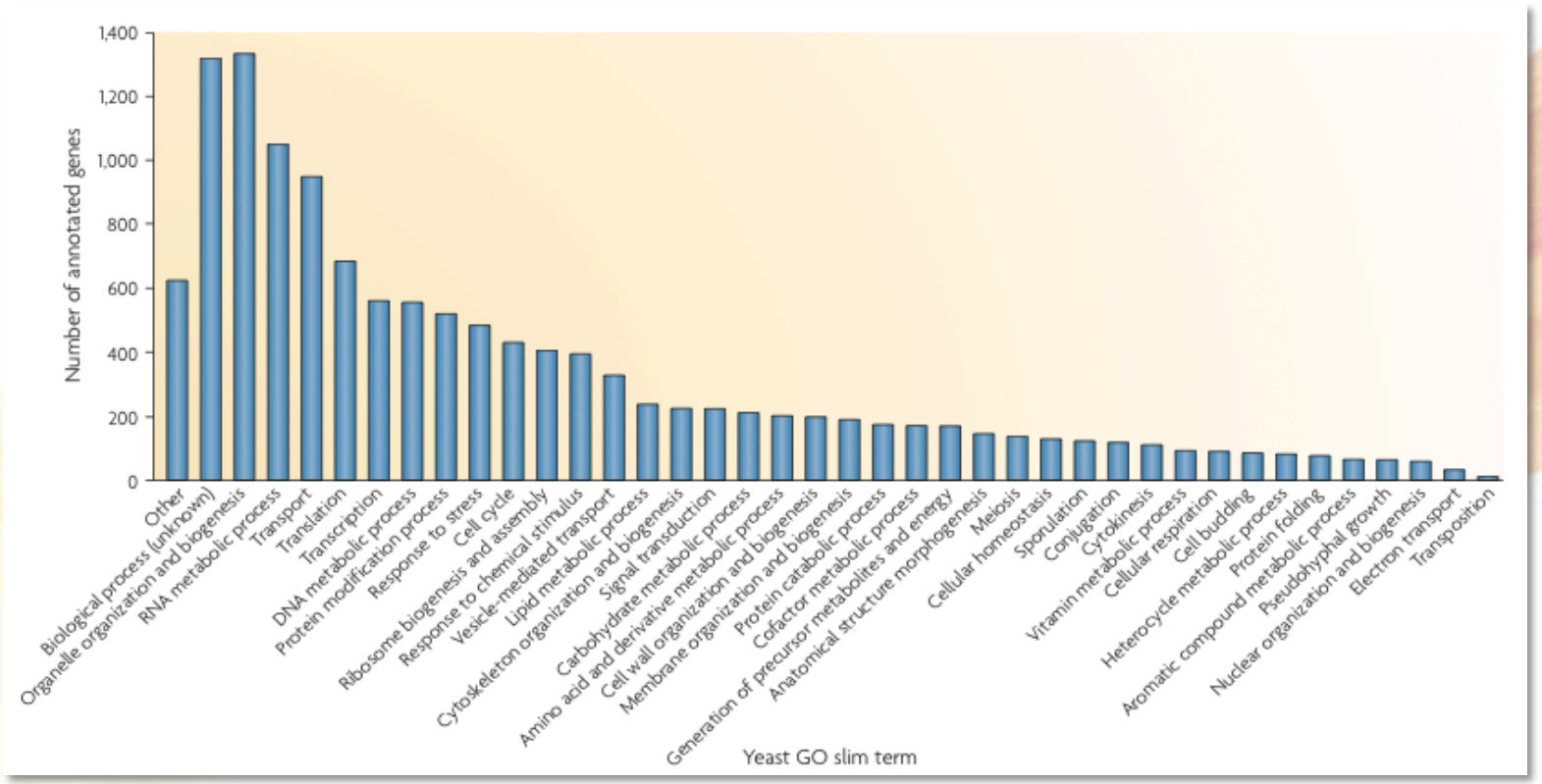


- Means genes can be grouped according to user-defined levels
- Allows broad overview of gene set or genome



4 - Bin Genome into Broad Biological Categories

- Using GO to bin the yeast genome into broad biological process categories



5 - As Annotation Pipeline

- Overview of Sequencing
- Sequence of genes/genome

ATGCTTCCTGATTTTGCCCTGGACTTCGCTTGTATAAATTCATTGCACC…

- Secondary Annotation - the functions of the genes

GO process: terrequinone A biosynthesis

GO function: methyltransferase activity

Enzyme Commission: 2.1.1.-

6 - Compile Data on Human Protein Coding Genes

- UniProtKB GO annotation program aims to provide high-quality Gene Ontology (GO) annotations to proteins in the UniProt Knowledgebase (UniProtKB)
- The assignment of GO terms to UniProt records is an integral part of UniProt biocuration
  - http://www.uniprot.org/help/biocuration
- UniProt manual and electronic GO annotations are supplemented with manual annotations supplied by external collaborating GO Consortium groups, to ensure a comprehensive GO annotation dataset is supplied to users
- UniProtKB-GOA is a member of the GO Consortium
- Go to:
  - GOA at ebi - http://www.ebi.ac.uk/GOA/
  - Select "Human" on the left
  - Download the GOA for Human
  - ftp://ftp.ebi.ac.uk/pub/databases/GO/goa/HUMAN/gene_association.goa_human.gz
  - GO
    - http://www.geneontology.org/ontology/obo_format_1_2/gene_ontology_ext.obo

Getting the GO Data

Get the GO Data

- Gene Ontology project
  - Major bioinformatics initiative
  - Aim of standardizing the representation of gene and gene product attributes across species and databases
- Provides a controlled vocabulary of terms
  - Terms, definitions and ontology structure
    - http://www.geneontology.org/GO.downloads.ontology.shtml
  - Gene association files data from GO Consortium members**
    - http://www.geneontology.org/GO.downloads.annotations.shtml
  - As well as tools to access and process this data
    - http://www.geneontology.org/GO.tools.shtml
- Get this data
  - http://www.geneontology.org/ontology/obo_format_1_2/gene_ontology_ext.obo
  - ftp://ftp.ebi.ac.uk/pub/databases/GO/goa/HUMAN/gene_association.goa_human.gz

** see next 2 slides

GO Current Annotations: Filtered Files

- These files are taxon-specific and reflect the work of specific projects
- All the files in this table have been filtered using the annotation file QC checks script

Species, Database	Gene Products Annotated	Annotations	Submission date *MM/DD/YYYY*	Download filtered files
*Agrobacterium tumefaciens*str. C58 PAMGO	82	248 (248 non-IEA)	9/24/2012	annotations [3.5 kb] README
Arabidopsis thaliana TAIR	30326	180225 (180215 non-IEA)	10/11/2013	annotations [6.1 mb] README
Aspergillus nidulans AspGD	47595	219175 (83799 non-IEA)	10/11/2013	annotations [2.5 mb] README
Comprehensive Microbial Resource [multispecies] JCVI	61681	154245 (154245 non-IEA)	9/13/2013	annotations [1.8 mb] README
Bos taurus GO Annotations @ EBI	19697	111414 (15232 non-IEA)	10/11/2013	annotations [1.5 mb] README
Caenorhabditis elegans WormBase	16715	103832 (60727 non-IEA)	10/11/2013	annotations [1.1 mb] README

http://www.geneontology.org/GO.current.annotations.shtml

GO Current Annotations: Unfiltered Files

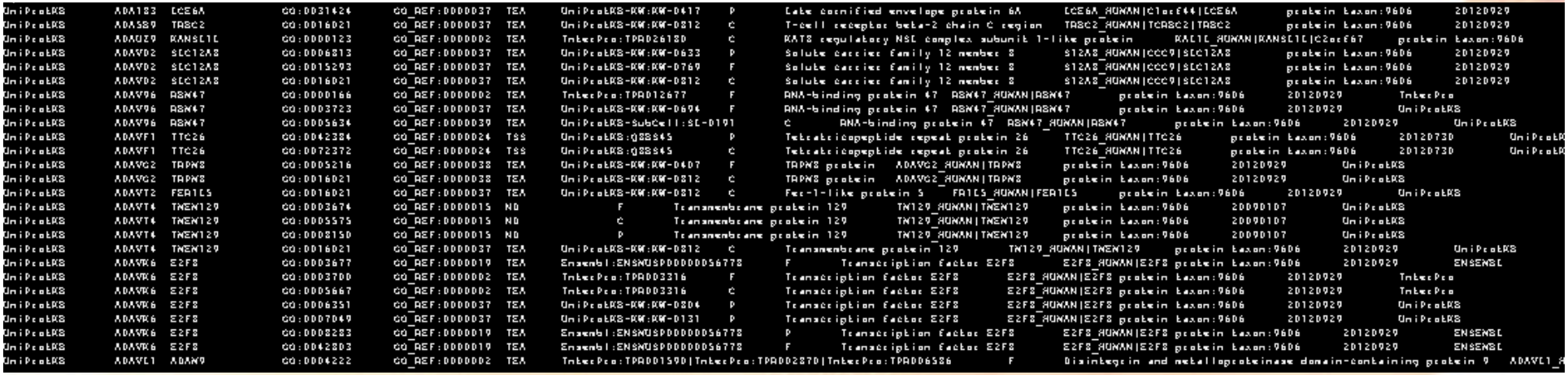
- These files have not been filtered with the annotation file QC checks script
- Most important difference between these files and the filtered files above is that gene products from certain taxa are *not* stripped out of the file

Species, Database	Gene Products Annotated	Annotations	Submission date *MM/DD/YYYY*	Download unfiltered files
Protein Data Bank [multispecies] GO Annotations @ EBI	176214	2538489 (687033 non-IEA)	10/2/2012	annotations [14.2 mb] README
Reactome [multispecies] CSHL & EBI	12934	97907 (97907 non-IEA)	8/30/2012	annotations [1.1 mb] README
UniProt [multispecies] GO Annotations @ EBI	17408324	110329676 (1145210 non-IEA)	10/2/2012	annotations [1.2 gb] README

http://www.geneontology.org/GO.current.annotations.shtml

The Gene Association File

- 17 Tab delimited columns



ftp://ftp.ebi.ac.uk/pub/databases/GO/goa/HUMAN/gene_association.goa_human.gz

Gene Association file

Column	Content	Required?	Cardinality	Example
1	DB	required	1	UniProtKB
2	DB Object ID	required	1	P12345
3	DB Object Symbol	required	1	PHO3
4	Qualifier	optional	0 or greater	**NOT**
5	GO ID	required	1	GO:0003993
6	DB:Reference (|DB:Reference)	required	1 or greater	PMID:2676709
7	Evidence Code	required	1	**IMP**
8	With (or) From	optional	0 or greater	GO:0000346
9	Aspect	required	1	F
10	DB Object Name	optional	0 or 1	Toll-like receptor 4
11	DB Object Synonym (|Synonym)	optional	0 or greater	hToll|Tollbooth
12	DB Object Type	required	1	protein
13	Taxon(|taxon)	required	1 or 2	taxon:9606
14	Date	required	1	20090118
15	Assigned By	required	1	SGD
16	Annotation Extension	optional	0 or greater	*part_of*(CL:0000576)
17	Gene Product Form ID	optional	0 or 1	UniProtKB:P12345-2

http://www.geneontology.org/GO.format.gaf-2_0.shtml

Gene Association file

Column	Code	Definition	Example
1	DB	Database from which annotated entry has been taken	UniProtKB
2	DB_Object_ID	unique identifier in the database for the item being annotated	O00165
3	DB_Object_Symbol	officially approved gene symbol will be used in this field when available	G6PC
4	Qualifier	flags that modify the interpretation of an annotation	NOT
5	GO ID	GO identifier for the term attributed to the DB_Object_ID	GO:0005634
6	DB:Reference	reference cited to support an annotation	PMID:9058808
7	Evidence	EXP, IMP, IC, IGI, IPI, ISS, IDA, IEP, IEA, TAS, NAS, NR, ND or RCA.	EXP
8	With	additional identifier to support annotations using certain evidence codes (including IEA, IPI, IGI, IMP, IC and ISS evidences).	UniProtKB:O00341
9	Aspect	P (biological process), F (molecular function) or C (cellular component). Example: P	P
10	DB_Object_Name	full UniProt protein name will be present here, if available from UniProtKB	subunit beta
11	Synonym	Alternative gene symbol(s), IPI identifier(s) and UniProtKB/Swiss-Prot identifiers are provided pipe-separated, if available from UniProtKB	IPI00706050
12	DB_Object_Type	What kind of entity is being annotated	protein
13	Taxon_ID	Identifier for the species being annotated	taxon:9606
14	Date	The date of last annotation update in the format 'YYYYMMDD'	20050101
15	Assigned_By	Attribute describing the source of the annotation	UniProtKB
16	Annotation_Extension	Contains cross references to other ontologies/databases that can be used to qualify or enhance the GO term applied in the annotation	occurs_in(GO:0009536)
17	Gene_Product_Form_ID	The unique identifier of a specific splice-form of the protein described in column 2	O43526-2

gene_ontology_ext.obo

- Current format used by the Gene Ontology Consortium (GOC) is called OBO v1.2
- extended version – gene_ontology_ext.obo

```
[Term]
id: GO:0000001
name: mitochondrion inheritance
namespace: biological_process
def: "The distribution of mitochondria, including the mitochondrial genome, into daughter cells after
mitosis or meiosis, mediated by interactions between mitochondria and the cytoskeleton." [GOC:mcc, PMI
D:10873824, PMID:11389764]
synonym: "mitochondrial inheritance" EXACT []
is_a: GO:0048308 ! organelle inheritance
is_a: GO:0048311 ! mitochondrion distribution

[Term]
id: GO:0000002
name: mitochondrial genome maintenance
namespace: biological_process
def: "The maintenance of the structure and integrity of the mitochondrial genome; includes replication
 and segregation of the mitochondrial chromosome." [GOC:ai, GOC:vw]
is_a: GO:0007005 ! mitochondrion organization

[Term]
id: GO:0000003
name: reproduction
namespace: biological_process
alt_id: GO:0019952
alt_id: GO:0050876
def: "The production by an organism of new individuals that contain some portion of their genetic mate
rial inherited from that organism." [GOC:go_curators, GOC:isa_complete, ISBN:0198506732]
subset: goslim_generic
subset: goslim_pir
subset: goslim_plant
subset: gosubset_prok
synonym: "reproductive physiological process" EXACT []
xref: Wikipedia:Reproduction
is_a: GO:0008150 ! biological_process
```

http://www.geneontology.org/ontology/README

Perl to the Rescue – Two Years Ago

GO	Hits	Name	Namespace
GO:0005515	**14508**	protein binding	molecular_function
GO:0005829	9114	***cytosol***	cellular_component
GO:0005634	8957	***nucleus***	cellular_component
GO:0005886	8678	***plasma membrane***	cellular_component
GO:0005737	**7579**	cytoplasm	cellular_component
GO:0016021	6855	integral to membrane	cellular_component
GO:0016020	**6144**	**membrane**	**cellular_component**
GO:0005654	5325	nucleoplasm	cellular_component
GO:0005576	4886	extracellular region	cellular_component
GO:0046872	4257	metal ion binding	molecular_function
GO:0008270	4154	zinc ion binding	molecular_function
GO:0005622	4065	intracellular	cellular_component
GO:0000166	3871	nucleotide binding	molecular_function
GO:0005524	3592	ATP binding	molecular_function
GO:0003677	2820	DNA binding	molecular_function
GO:0006355	2502	regulation of transcription, DNA-dependent	biological_process
GO:0005739	2302	mitochondrion	cellular_component
GO:0004872	2294	receptor activity	molecular_function
GO:0005488	2087	binding	molecular_function
GO:0003700	1931	sequence-specific DNA binding transcription factor activity	molecular_function

With those two files you can produce something like above

Perl to the Rescue – Last Year

GO	Hits	name	namespace
GO:0005515	**21303**	protein binding	molecular_function
GO:0005829	10776	***cytosol***	cellular_component
GO:0005886	10020	***plasma membrane***	cellular_component
GO:0005634	9216	***nucleus***	cellular_component
GO:0005737	**7018**	cytoplasm	cellular_component
GO:0016021	6656	integral to membrane	cellular_component
GO:0005654	5691	nucleoplasm	cellular_component
GO:0005576	5316	extracellular region	cellular_component
GO:0008270	4352	zinc ion binding	molecular_function
GO:0005524	3789	ATP binding	molecular_function
GO:0005730	2814	nucleolus	cellular_component
GO:0005622	2635	intracellular	cellular_component
GO:0003677	2619	DNA binding	molecular_function
GO:0005739	2134	mitochondrion	cellular_component
GO:0006351	2119	transcription, DNA-dependent	biological_process
GO:0006355	2044	regulation of transcription, DNA-dependent	biological_process
GO:0003676	1915	nucleic acid binding	molecular_function
GO:0003700	1899	sequence-specific DNA binding transcription factor activity	molecular_function
GO:0016020	**1857**	**membrane**	**cellular_component**
GO:0005509	1560	calcium ion binding	molecular_function

With those two files you can produce something like above

Conclusion of GO

- GO produces sets of:
  - Explicitly defined, structured vocabularies that
  - That describe:
    - biological processes
    - molecular functions
    - cellular components
  - Of gene products in both a computer- and human-readable manner
- **Available annotation for a given organism might affect results and conclusions**
- Care should be taken when choosing an analysis method - Might be essential to **include or exclude** certain types of annotations for certain types of analysis
- The GO is a tool becoming increasingly powerful for data analysis and functional predictions as the **ontologies and annotations continue to evolve**

For Thursday

- Read:
  - **OPTIONAL:** The Alu RNAs embedded in 5' and 3' UTRs of human mRNAs
    - http://www.unige.ch/sciences/biologie/bicel/Strub/researchAlu.html
  - Use and Misuse of the Gene Ontology Annotations course website
    - http://155.33.203.128/teaching/BIOL6308-Fall2013/local/Literature/UseAndMisuseGoAnnotations.pdf
  - Prepare for a quiz