BioSci D145 Lecture #6

- Bruce Blumberg (blumberg@uci.edu)
  - 4103 Nat Sci 2 - office hours Tu, Th 3:30-5:00 (or by appointment)
  - phone 824-8573

- TA – Angela Kuo (akuo4@uci.edu)
  - 4311 Nat Sci 2– office hours W 10-12
  - Phone 824-6873

- Updated lectures (with answers) will be posted after lecture
  - http://blumberg-lab.bio.uci.edu/biod145-w20120

Term paper requirements and scoring

- Actual paper – 5 pages single spaced 1" margins (references not included).

- PLEASE DO NOT WRITE AN ABSTRACT

- **Specific aims – 2 points** (this should be about 3/4 to one page)
  - Write a paragraph introducing the topic, state why it is important and what are the gaps in knowledge that you will address.
  - For example,
    - something is a problem.
    - we know this, and we know this and we know this, etc
    - BUT we don't know that
    - A big question in the field that will help us address that is…
    - We hypothesize that…
    - We propose the following specific aims to test this hypothesis

Term paper requirements and scoring

- Actual paper – 5 pages single spaced 1" margins (references not included).

- PLEASE DO NOT WRITE AN ABSTRACT

- **Specific aims – 2 points** (this should be about 3/4 to one page)
  - Enumerate 2-3 specific aims in the form of questions that test your hypothesis. Use 1-2 sentences to state how you will answer the question
  - Finish with a paragraph entitled "Impact" - describing what will be the impact of the work you are proposing.
    - That is, why is it important and what large benefit will derive from what you are proposing
    - How will it move the field forward
    - Why is this important?
  - This is a critically important part of a grant application. You have to convince the reviewer that your work is important and worth funding.
    - Otherwise they will fund someone else who has taken the time to make the case for why it is important and will move the field forward.
  - http://blumberg-lab.bio.uci.edu/biod145-w2020/specific_aims.pdf

Term paper requirements and scoring

- **Background and Significance – 3 points** (about 1.5-2.5 pages)
  - Briefly summarize what is known about the problem.
    - Not a comprehensive review, just a summary of the important points.
    - Expand on what you wrote in specific aims
  - Succinctly state what is not known and why it is important that this research be done
    - Address knowledge gaps
    - Are you addressing something controversial?
      - talk about the controversy and why your work will address it directly.
  - In about one paragraph at the end discuss the significance of your proposed research and how it is innovative
    - Significance = why will accomplishing it benefit the research community and world at large if you are successful
    - Innovation – what is particularly innovative about what you are proposing with respect to methodology, questions asked, etc.

- http://blumberg-lab.bio.uci.edu/biod145-w2020/sigificance_innovation.pdf

Term paper requirements and scoring

- **Research plan – 4 points** (about 2.5-3 pages)
  - In a short paragraph, state what you will do and why it is important. (I know it seems repetitive by now, but reviewers are busy and will be skimming your grant. You need to hit them over the head a few times before they will get your point).
  - Restate each specific aim from the Specific aims section (one by one)
  - describe what you will do to address the aim
    - Break into subaims as appropriate
    - State the hypothesis to be tested in each
    - Explain the rationale
    - Describe briefly what approach you will take
    - Discuss what you expect to find
    - Point out any possible problems and alternative approaches
  - I am mostly concerned with your hypothesis and rationale here.
  - Not an all-encompassing proposal – 4-5 years by a small team (e.g., your PhD thesis research)
- http://blumberg-lab.bio.uci.edu/biod145-w2020/Intro_research_plan.pdf

Comparative genomics

- Study of similarities and differences between genome structure and organization
  - How many genes? Chromosomes?
  - Genome duplications
  - Gene loss
- Driving forces
  - Understanding evolution in molecular terms
  - Sequence annotation and function identification
    - Sequences with important functions often evolutionarily conserved
- Orthology vs paralogy
  - Homolog – descended from a common ancestor (Hox genes)
  - Orthologs - homologous genes in different organisms that encode proteins with the same function and which have evolved by direct vertical descent (frog and human Hoxa-1)
  - Paralogs – homologous genes that encode proteins with related but non-identical functions (Hoxa-1, Hoxb-1, Hoxd-1)
  - Homeolog - Polyploid copies derived from duplication or mating, e.g., duplicated genes in tetraploid organisms (Hox-a1a, Hox-a1b

Comparative genomics (contd)

- Functional equivalency does not require homology, sequence similarity or even 3D structure
  - Same chemical reaction can be catalyzed by totally unrelated enzymes
  - **Non-orthologous gene displacement** – when non-orthologous genes encode the same essential cellular function
    - Better term would be **analogous** gene
    - Convergent evolution also sometimes used

Table 1. Dissimilar Enzymes Catalyzing the Same Biochemical Reactions[a]

Enzyme activity (EC No.)	Taxonomic representation[b] bacteria	archaea	eukaryotes	PDB entry	Structural folds[c]
Alcohol:NADP dehydrogenase	**ADH_CLOBE**	ADH3_SULSO	**ADH1_ENTHI**	**1DEH**	different
(EC 1.1.1.2)	DHSO_BACSU	—	**ALDX_HUMAN**	**2ALR**	
Formate dehydrogenase	**FDHF_ECOLI**	**FDHA_METFO**	—	**1FDI**	different
(EC 1.2.1.2)	**FDH_PSESR**	A64427	**FDH_NEUCR**	**2NAD**	
Dihydrofolate reductase	**DYRA_ECOLI**	**DYR_HALVO**	**DYR_HUMAN**	**1DHF**	different
(EC 1.5.1.3)	**DYR2_ECOLI**	—	—	**1VIE**	
Peroxidase	—	—	**PERM_HUMAN**	**1MHL**	same,
(EC 1.11.1.7)	—	—	**PER1_ARAHY**	**1ARV**	RMSD = 4.8
Chloroperoxidase	**PRXC_PSEPY**	—	—	1BRO	different
(EC 1.11.1.10)	—	—	**PRXC_CALFU**	**1CPO**	different
	—	—	**PRXC_CURIN**	**1VNC**	
Superoxide dismutase	**SODC_ECOLI**	—	**SODC_HUMAN**	**1SPD**	different
(EC 1.15.1.1)	**SODF_ECOLI**	**SODF_SULAC**	**SODM_HUMAN**	**1ABM**	
Protein-tyrosine phosphatase	**PTPA_STRCO**	—	**PPAC_BOVIN**	**1PHR**	different
(EC 3.1.3.48)	**YOPH_YEREN**	—	**PTN1_HUMAN**	**2HNP**	
Cellulase	**GUNA_CLOCE**	—	**GUNB_NEOPA**	**1EDG**	different
(EC 3.2.1.4)	**GUND_CLOTM**	—	**GUN_PHAVU**	**1CLC**	different
	—	—	**GUN1_TRIRE**	**1CEL**	
Xylanase	**XYNA_STRLI**	—	**S43846**	**1XAS**	different
(EC 3.2.1.8)	**XYNA_BACCI**	—	**XYN2_TRIRE**	**1XNB**	
Chitinase	**CHIA_SERMA**	—	**CHIT_BRUMA**	**1CTN**	different
(EC 3.2.1.14)	YE15_HAEIN	—	**CHI1_ORYSA**	**2BAA**	
β-Galactosidase	**BGAL_ECOLI**	—	**BGAL_KLULA**	**1BGL**	different
(EC 3.2.1.23)	BGLA_THEMA	**BGAM_SULSO**	BGLC_MAIZE	**1GOW**	
Lichenase	**GUB_BACLI**	—	YG46_YEAST	**1GBG**	different
(EC 3.2.1.73)	**GUB_BACCI**	—		**1CEM**	different
	—	—	**GUB2_HORVU**	**1GHR**	
β-Lactamase	**AMPC_ENTCL**	—	—	**2BLT**	different
(EC 3.5.2.6)	**BLAB_BACFR**	—	—	**1ZNB**	
Fructose 1,6-bisphosphate	**ALF_ECOLI**	—	**ALF_YEAST**	**1DOS**	same,
aldolase (EC 4.1.2.13)	**ALF_STACA**	—	**ALFA_HUMAN**	**1FBA**	RMSD = 3.4
Carbonic anhydrase	CCMM_SYNP7	**CAH_METTE**	—	**1THJ**	different
(EC 4.2.1.1)	—	—	**CAH1_HUMAN**	**2CBA**	
Peptidyl-prolyl isomerase	**FKBX_ECOLI**	FKB1_METJA	**FKBP_HUMAN**	**1FKD**	different
(EC 5.2.1.8)	**CYPB_ECOLI**	—	**CYPB_HUMAN**	**2CPL**	
Chorismate mutase	**PHEA_ECOLI**	Y246_METJA	**CHMU_YEAST**	**1ECM**	different
(EC 5.4.99.5)	**CHMU_BACSU**	—	—	**1COM**	
DNA topoisomerase I	**TOP1_ECOLI**	**TOPG_SULAC**	**TOP3_YEAST**	**1ECL**	different
(EC 5.99.1.2)	—	—	**TOP1_YEAST**	**1OIS**	

[a]The full version of the table, including homologs of the enzymes found in each of the sequenced genomes, is available as a WWW supplement at http://ncbi.nlm.nih.gov/Complete_Genomes.
[b]The proteins are listed under their SwissProt, GenBank, or Protein Data Base identifiers. The names of enzymes with experimentally demonstrated activity, shown in the first column, are in boldface type; the dash indicates absence of homologs in any of the sequenced genomes.
[c]The data are from SCOP [http://scop.mrc-lmb.cam.ac.uk/scop (Hubbard et al. 1997)] and FSSP [http://www2.ebi.ac.uk./dali/fssp/fssp.html (Holm and Sander 1996a)] databases. RMSD of superimposed Cα atoms in the structural alignment of the two isoforms is from the FSSP database (Holm and Sander 1996a).

Comparative genomics (contd)

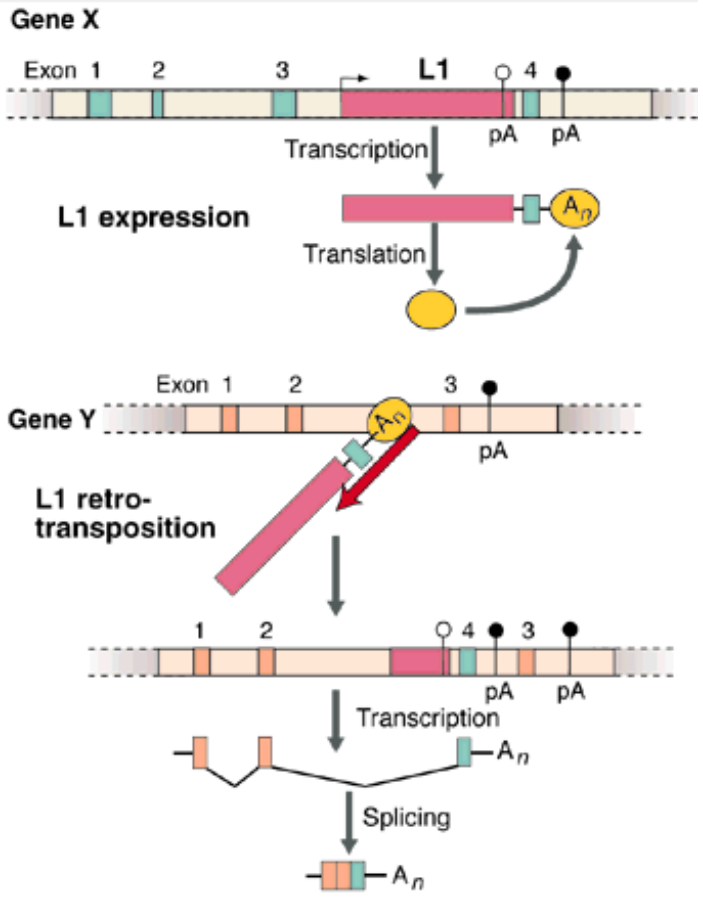
- Genes with very different functions can be related
  - 3-D structure may indicate that proteins are related (evolved from the same ancestral protein) but sequence identity too low to detect
    - Expected when genes diverge from a distant common ancestor
    - < 20% amino acid sequence identity too little to establish homology (although proteins may be homologous)
  - For example
    - 3-D structures of
      - D-alanine ligase
      - Glutathione synthetase
      - ATP-binding domains of
        - » Carbamoyl phosphate sythetase
        - » Succinyl-CoA synthetase
    - Are all so similar in 3D structure that homology is not in doubt but sequence comparisons do not detect homology

- Why should we care whether genes are related or not?

  Essential for understanding how evolution works at the molecular level

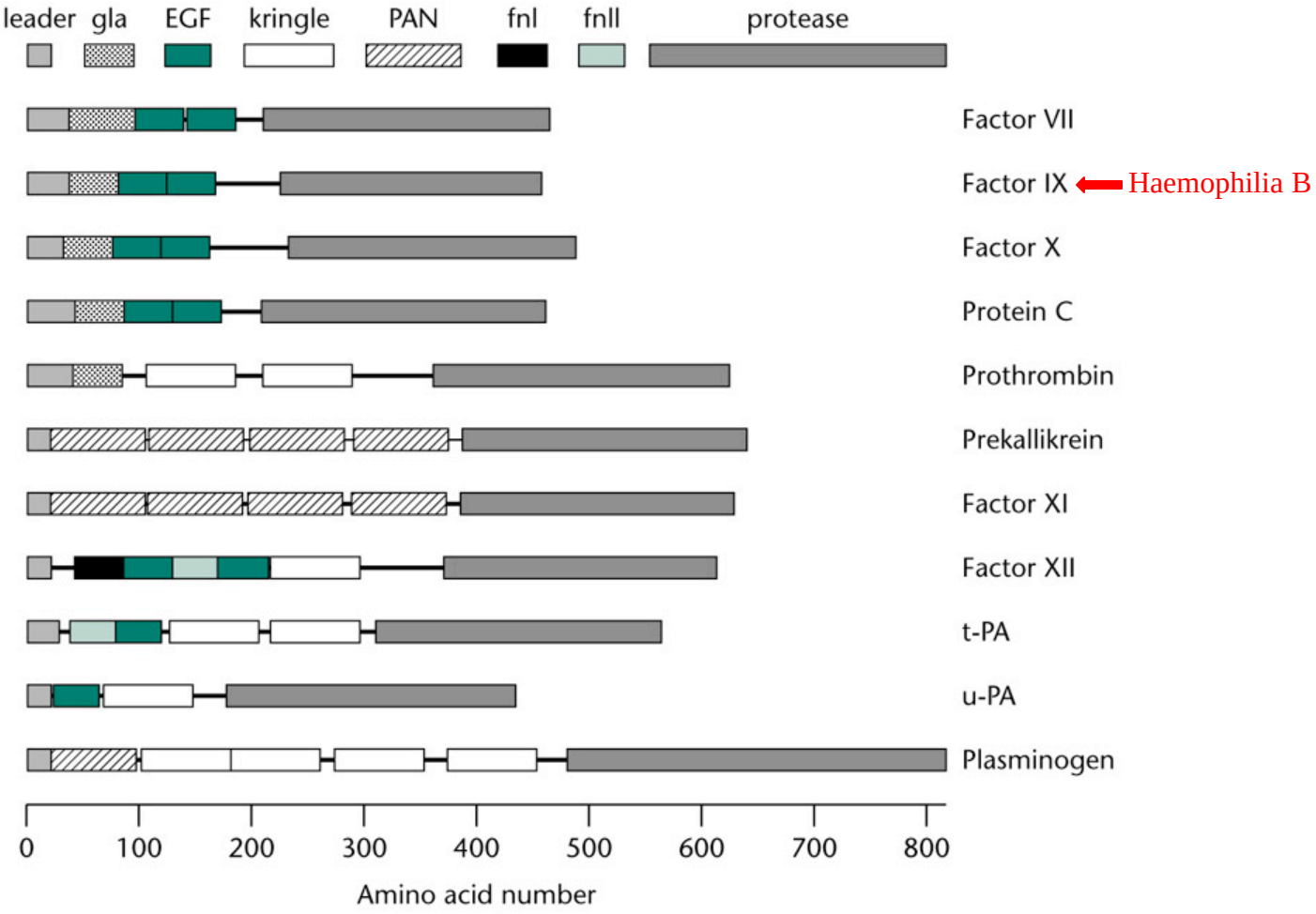
Comparative genomics (contd)

- Protein evolution
  - Observation – many proteins composed of discrete domains
  - Observation – many proteins have multiple domains shared with other proteins
  - Conclusion – domain shuffling must have occurred during evolution
  - Some correlation between exons and protein domains
    - Protein domains tend to be encoded in 1 or two exons
    - New combinations of protein domains can be created by recombination
      - LINEs
      - Between repetitive elements in introns
    - **Exon shuffling** – process of transferring exons (and hence functional domains) between proteins



Comparative genomics (contd)

- Protein evolution (contd)
  - Haemostatic (aka blood clotting) proteins as an exon shuffling paradigm
    - Family of proteases that are activated by proteolysis
    - Protein domains show strong correlation with exons



Comparative genomics (contd)

- Protein evolution (contd)
  - **What is horizontal gene transfer** – transfer of genes or protein domains across unrelated species
    - Frequently identifiable by different patterns of codon usage from other genes, particularly ribosomal proteins
    - Fairly rare with eukaryotes
    - Happens in prokaryotes all the time – Examples?
      - e.g., transfer of antibiotic resistance among bacteria
      - Plasmid exchange, phage infections and transfer
      - Often associated with pathogenicity
        - » Pathogenic variants of bacteria frequently have lots of inserted DNA
        - » e.g., *E. coli* H0157 has 800 kb more than lab strains of *E. coli,* much of which is virulence factors, prophages and prophage like elements
      - What does this suggest about nature of virulence?

        Virulence is acquired, i.e, transferred from one organism to another

Comparative genomics (contd)

- Is there a **minimal genome**? How would you define "minimal genome"?
  - Encoding the essential set of proteins required for life?
  - Compare genomes of archebacteria, eubacteria and yeast
    - Issues with how genes are classified but a reasonably good approximation can be made
    - Can identify 322 clusters of orthologous groups required for all key biosynthetic pathways that might be required in free-living organisms
      - But remember about non-orthologous gene displacements!
- Some lessons from bacterial genomics
  - Nearly half of ORFs are of unknown function
  - About 25% of all ORFs are unique to a particular species!
    - Suggests that many new protein families remain to be discovered
    - Many new functions may be uncovered
  - Periodic re-evaluation of sequenced genomes is useful
    - Compare with newly acquired data
      - Often find additional ORFs and genes
  - Much conservation of gene position
    - Same genes found in many genomes at same positions (good for evolutionary studies

Comparative genomics (contd)

- What do we get from comparative genomics?
  - Powerful new tools to identify conserved sequences
    - important regulatory elements
    - Unidentified genes
    - Features (promoters, splice sites, etc)
  - Important information about genome evolution
    - Where did related genes originate?
    - When did genome duplications arise?
    - What is the history of life on earth?
      - And by implication, life elsewhere
    - What is the genetic diversity in wild populations
      - Environmental shotgun sequencing
  - Information required to identify gene function
    - Protein sequence and structure comparisons

Construction of cDNA libraries

- What is a cDNA library?
  - Collection of DNA copies representing the expressed mRNA population of a cell, tissue, organ or embryo

- What are they good for?
  - Identifying and isolating expressed mRNAs
  - functional identification of gene products
  - cataloging expression patterns for a particular tissue
    - EST sequencing and microarray analysis
  - Mapping gene boundaries
    - Promoters
    - Alternative splicing

Determinants of library quality

- What constitutes a full-length cDNA?
  - Strictly, it is an exact copy of the mRNA
  - full-length protein coding sequence considered acceptable for most purposes
- mRNA
  - full-length, capped mRNAs are critical to making full-length libraries
  - cytoplasmic mRNAs are best – WHY?

    They are processed, i.e., introns removed and poly A is added
- 1st strand synthesis
  - complete first strand needs to be synthesized
  - issues about enzymes
- 2nd strand synthesis
  - thought to be less difficult than 1st strand (probably not)
- choice of vector
  - plasmids are best for EST sequencing and functional analysis
  - phages are best for manual screening

cDNA synthesis

- Scheme
  - mRNA is isolated from source of interest
  - 1-10 µg are denatured and annealed to primer containing $d(T)_nV$
    - To minimize length of poly A tail in libraries for sequencing
  - reverse transcriptase copies mRNA into cDNA
  - DNA polymerase I and Rnase H convert remaining mRNA into DNA
  - cDNA is rendered blunt ended
  - linkers or adapters are added for cloning
  - cDNA is ligated into a suitable vector
  - vector is introduced into bacteria

- Caveats
  - there is lots of bad information out there
    - much is derived from vendors who want to increase sales of their enzymes or kits
  - all manufacturers do not make equal quality enzymes
  - most kits are optimized for speed at the expense of quality
  - small points can make a big difference in the final outcome

Functional Genomics - The challenge: Many new genes of unknown function

- Where/when are they expressed?
  - Known genes (e.g. from genome projects)
    - Gene chips (Affymetrix) and microarrays (Oligo, cDNA, protein)
  - Novel genes
    - Expression profiling
      - Genomic tiling microarrays
      - SAGE and related approaches
      - Massively parallel sequencing (RNA-Seq) (Owen)
      - Single cell RNA-seq (sc-RNA-seq) (Morrison, Cheng)
- Personal 'omic approaches (Chen)
- Which genes regulate what other genes? (Flyamer, Buenrostro, Argelaget)
- What is the phenotype of loss-of-function? (week 8 papers)
  - Genome wide CRISPR editing (Anzalone)
  - CRISPR/Cas approaches to gene regulation (Gilbert)
  - Genome wide synthetic lethal screens (Luo)
- Detecting protein-protein interactions (week 9 papers)
- Metabolome & microbiome (week 10 papers)

Routes to gene identification

- Genome sequences are minimally useful without annotation
  - Annotation = description, biological information
    - Functional annotation – information on the function
    - Structural annotation – identification of genes, sequence elements
  - Much annotation is done automatically today
    - Via sequence comparisons with various databases
      - Gene sequences
      - ESTs
    - Algorithms predict promoters, splicing, polyadenylation sites and, most importantly ORFs
    - ORFs – open reading frames are putative proteins
      - Algorithms miss in both directions
      - Source of much disagreement

- Field of bioinformatics has grown to encompass many types of analysis related to gene function
  - www.igb.uci.edu

How are genes identified?

- Random
  - EST sequencing, select interesting looking gene
  - Large scale expression analysis
    - http://xenopus.nibb.ac.jp/
- From protein sequences
  - Antibody screening
  - Reverse translate and oligo screen
- Functional cloning
  - Finding a gene by using a functional assay
- Positional cloning
  - Find a gene by where it is located, what it is near
- By similarity to other sequences
  - Gene family
  - Cross-species
  - Computer based equivalents
- Bioinformatic analysis that relates back to functional or positional cloning

How are genes identified? (contd)

- Ways to identify genes in regions
  - Cross-species hybridization
    - Probe another species with this genomic region
    - coding sequences are conserved -> should see hybridization where genes are
    - What do you think are limitations to this approach?
      - Species must be sufficiently different to reduce "noise" from overall sequence conservation
        - » mouse vs human probably not great
        - » Human vs frog or fish probably good
      - Must be sufficiently similar for genes to be conserved
        - » Human vs frog or fish probably good
        - » Humans vs yeast only good for common genes
      - Target species region needs to be well characterized
    - Computer parallels – compare sequence to be annotated with annotated sequence from a different organism
      - e.g., human *with Drosophila*
      - Unknown bacterium with *E. coli*, etc.

How are genes identified? (contd)

- Ways to identify genes in regions (contd)
  - Hybridization to known genes or coding materials
    - What are some examples?
      - Hybridize to mRNA (Northern blots)
      - Hybridize to cDNA libraries (must be right tissue, cell or stage)
      - Hybridization to genomic tiling microarrays
      - Capture cDNAs or mRNAs from solution
    - Computer based parallels
      - Compare with expressed sequences from other species
      - Compare with ESTs

How are genes identified? (contd)

- Ways to identify genes in regions (contd)
  - Identify features found in typical promoters
    - What are promoters?

      Regions 5' to a gene that are required for expression
    - CpG islands – regions in eukaryotic genes that are hypomethylated
      - Undermethylated - methylation of promoter DNA typically inhibits gene expression
      - Digest with enzymes that have CG in recognition site that would be inhibited if methylated, e.g., SacII CCGCGG, run gel to check
        - » If nonmethylated (expressed) enzyme will cut, region will be hypersensitive, get chopped up.
        - » If methylated (not expressed) enzyme will not cut and region will not get digested
    - DNAse I hypersensitive (or MNase or MPE)
      - Similar principle – transcriptionally active DNA is "open"
      - If open, it is more sensitive to DNAse I than non-active DNA
      - Test by digestion and gel electrophoresis
    - Assay for transposase accessible chromatin (ATAC)-seq
      - Open chromatin is accessible to transposase

How are genes identified? (contd)

- The problem with all of these methods is that experiments are required
  - What do we do when sequences are coming in at the rate of tens of gigabases/month?
  - Need large-scale, robust, computerized methods to identify genes and annotate sequences!

- All bioinformatics depends on databases
  - UCI bioinformatics has some unique databases (e.g., fuzzy PubMed)
    - http://www.igb.uci.edu/tools/databases.html

- Three major databases of sequences (automatically duplicated)
  - GENBANK http://www.ncbi.nlm.nih.gov/Genbank/index.html
  - DNA Databank of Japan http://www.ddbj.nig.ac.jp/
  - European Molecular Biology Laboratory (EMBL) http://www.ebi.ac.uk/embl/index.html

Genome annotation – how to identify genes?

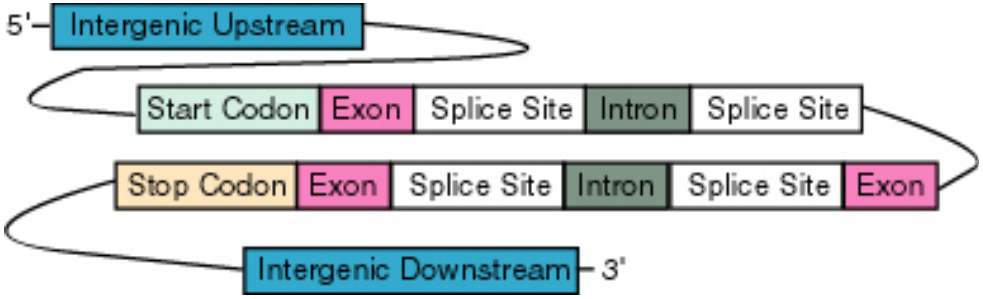
- Gene identification/prediction is important but difficult
  - Large variety of methods and algorithms to predict exons
  - To identify genes must first identify open reading frames (ORFs)
    - When dealing with cDNAs – look for regions that code for proteins
      - Do all genes code for proteins? Depends on definition of "gene"
      - Correct reading frame for a sequence is assumed to be largest with no stop codons (TGA, TAA, TAG)
    - Lots of tricks can be employed
      - Codon frequency for an organism
        - » Coding sequences follow codon usage
        - » Noncoding sequences do not, often have lots of stop codons
      - Consensus sites
        - » Kozak translational initiation CCRCCATGG
    - What is a very important consideration when searching sequences to predict ORFs?
      - Sequence must be accurate
        - » Incorrect base calls are troublesome
        - » But indels (insertions or deletions) are disastrous

Genome annotation – how to identify genes (contd)?

- Computer based gene prediction methods
  - Two major methods are in use
    - Homology searches
      - Compare with other known sequences
    - *Ab initio* (from the beginning) prediction
      - Use algorithms to recognize common features and predict genes
        - » Promoters
        - » Splice sites
        - » Polyadenylation sites
        - » ORFs
  - Generally, microbial genomes are much easier to annotate – WHY?

    Smaller - no or few introns
    - Simply identify ORFS > 300 bp (100 amino acids)
      - Works very well
      - But can miss small coding sequences
      - Must run on both strands because there are shadow genes (overlap on two strands)
    - Using a variety of programs, can predict genes in bacterial genomes
      - Venter Sargasso sea paper

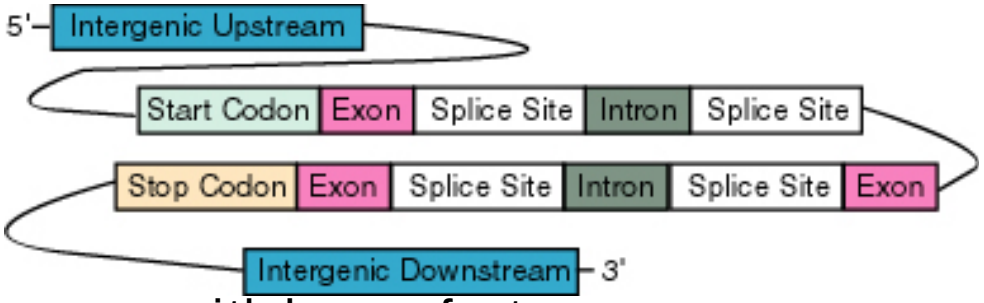
Genome annotation – how to identify genes (contd)?



- Computer based gene prediction methods (contd)
  - Huge variety of programs available
    - Neural networks – attempt to model learning process
      - Build decision trees, use probabilistic reasoning
    - Rule-based systems
      - Rules often not clear
      - Have trouble with exceptions
    - Hidden Markov models
      - Break sequences down into small units based on statistical analysis of composition
        - » Hexamers appear to be optimal size to search
      - Classify sequences into types or "states"
      - Identify transitions between states
      - Very useful for large number of purposes

Genome annotation – how to identify genes (contd)?

- Computer based gene prediction methods (contd)
  - Training sets are used to "teach" programs how to solve problems
    - Training set is actual data – genes with known features
  - Programs use training sets to classify new data
    - Neural networks use training data to build decision trees
    - Rule-based systems use training data to generate rules
    - HMM build table of probabilities for states and transitions
  - Pierre Baldi in IGB is UCI expert in machine learning



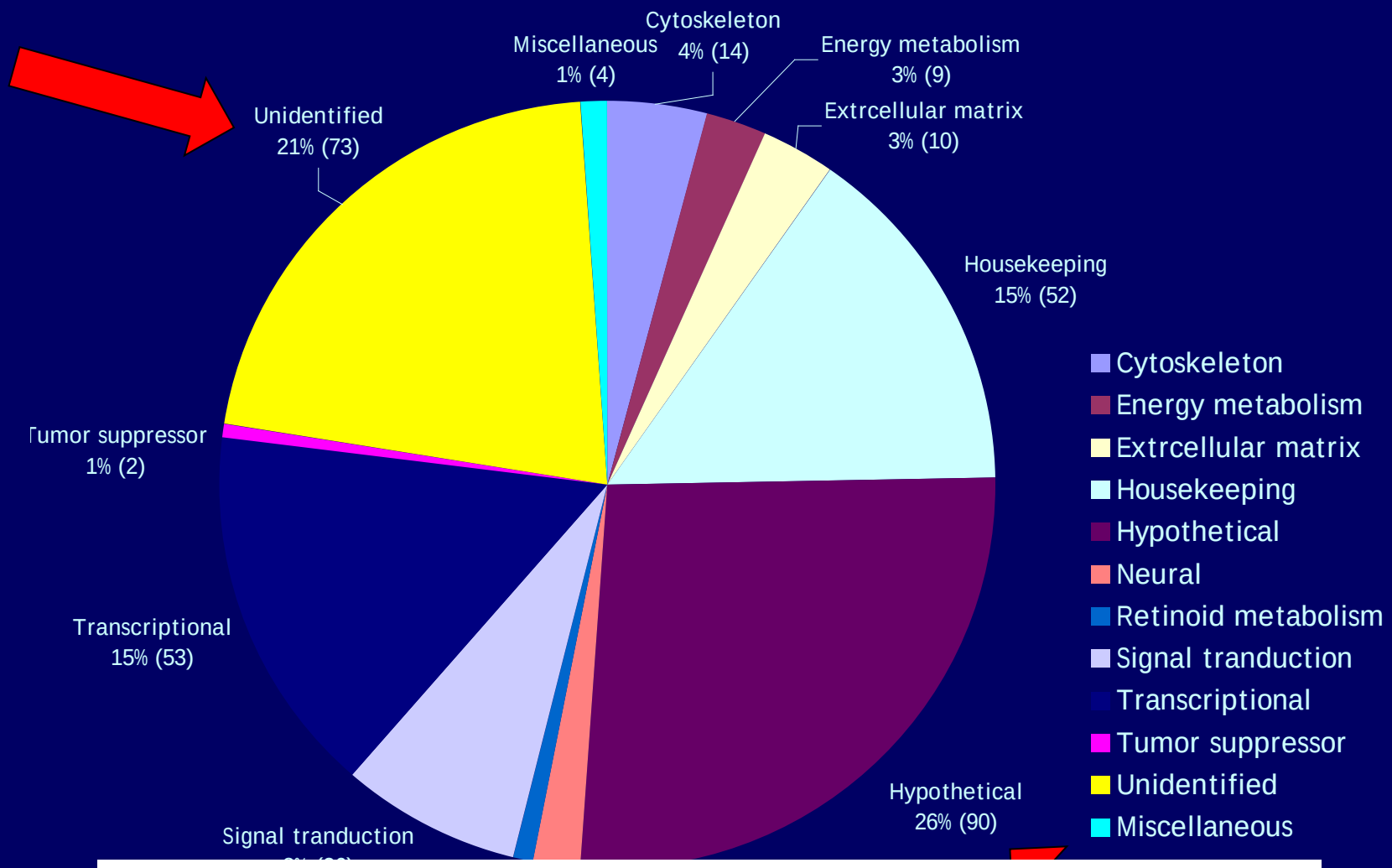
- How well do gene predicting programs work?
  - Extremely well on bacterial genomes
  - Fairly well on simply eukaryotic genomes
  - Variable on complex genomes
  - Rule of thumb – use a group of programs and look for areas of agreement among them
  - The best current programs combine *ab initio* predictions with similarity data to define a probability model

Identification of gene function

- You have identified a gene – what is its function?
  - Always look for similarity to known sequences
    - Swiss-prot is fairly well annotated
    - GENBANK translated database is most complete
    - BLAST is tool to use
    - Amino acid searches more sensitive than nucleotide searches
      - Because identical amino acid sequences might only be 67% identical at nucleotide level
  - What might you find?
    - Match may predict biochemical and physiological function
      - e.g., a known enzyme from another organism
    - Match may predict biochemical function only
      - e.g., a kinase
    - Match a gene from another organism with no known function
      - May match ESTs or ORFs from other organisms
    - Match a known gene with partly characterized function
      - Search leads to clarification of function
    - Might not match anything at all
      - Expect this will happen less and less

Up-regulated by TTNPB

(> 1.5 Fold, p < 0.01, n=334)

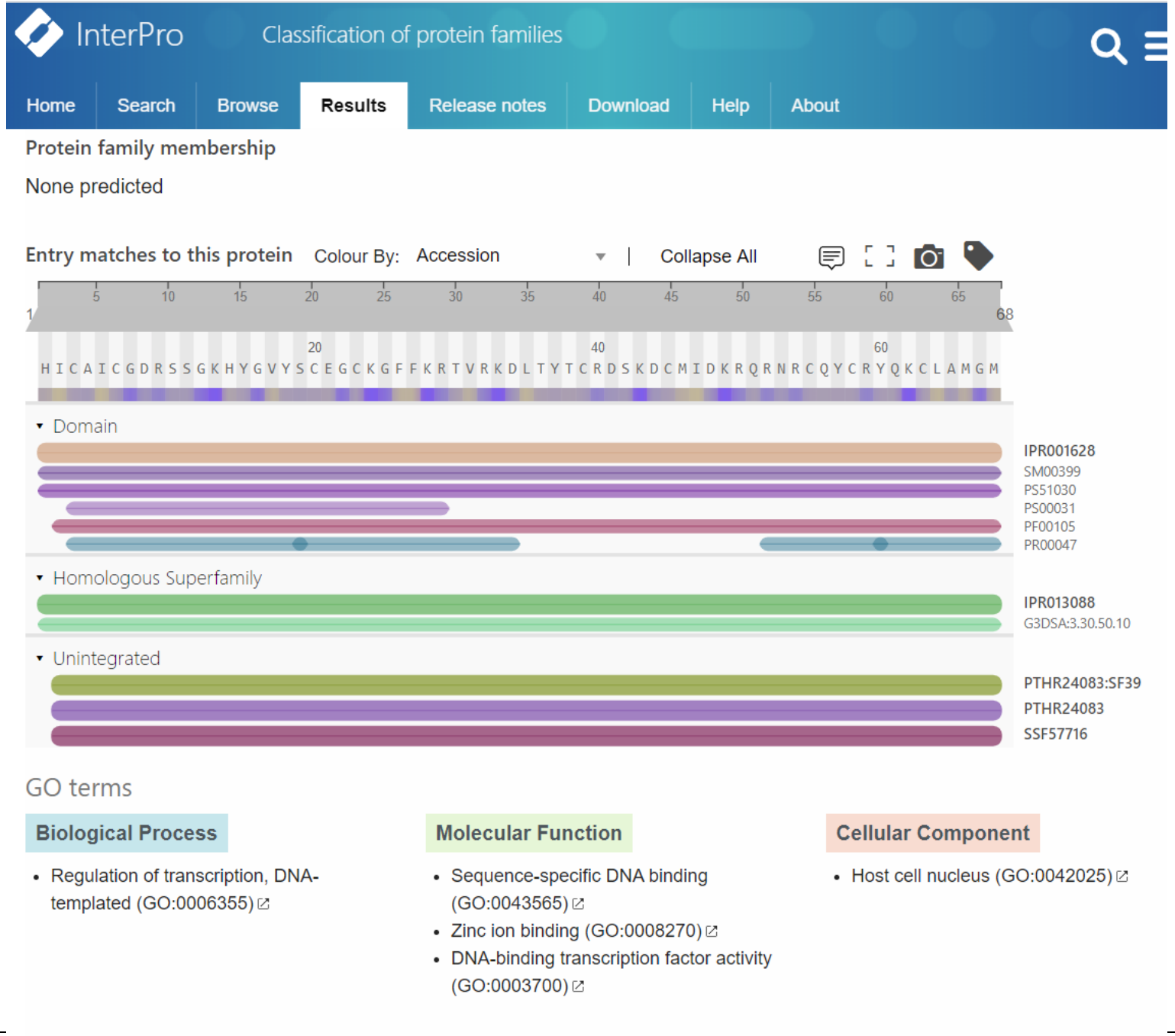


- Original study done in 2005
- Re-annotated in 2016, >20% still unidentified

Identification of gene function (contd)

- You have identified a gene – what is its function? (contd)
  - Does the sequence contain an obvious functional motif?
    - Homeobox or other consensus DNA binding domain?
    - Kinase domain?
    - Serine protease, etc.
  - InterPro database allows one to compare a protein sequence with whole family of structural databases
    - http://www.ebi.ac.uk/interpro/

      HICAICGDRSSGKHYGVYSCEGCKGFFKRTVRKDLTYTCRDSKDCMIDKRQRN RCQYCRYQKCLAMGM
    - https://www.ebi.ac.uk/interpro/result/InterProScan/iprscan5-R20200211-182400-0090-34212976-p1m/
    - Other sorts of similarity searches
    - Identify protein secondary structure motifs
      - Alpha helix, beta sheets, hydrophobicity
      - Amphipathic helices
      - Overall polarity of sequences
      - Not used much



GO terms

Biological Process

- Regulation of transcription, DNA-templated (GO:0006355)

Molecular Function

- Sequence-specific DNA binding (GO:0043565)
- Zinc ion binding (GO:0008270)
- DNA-binding transcription factor activity (GO:0003700)

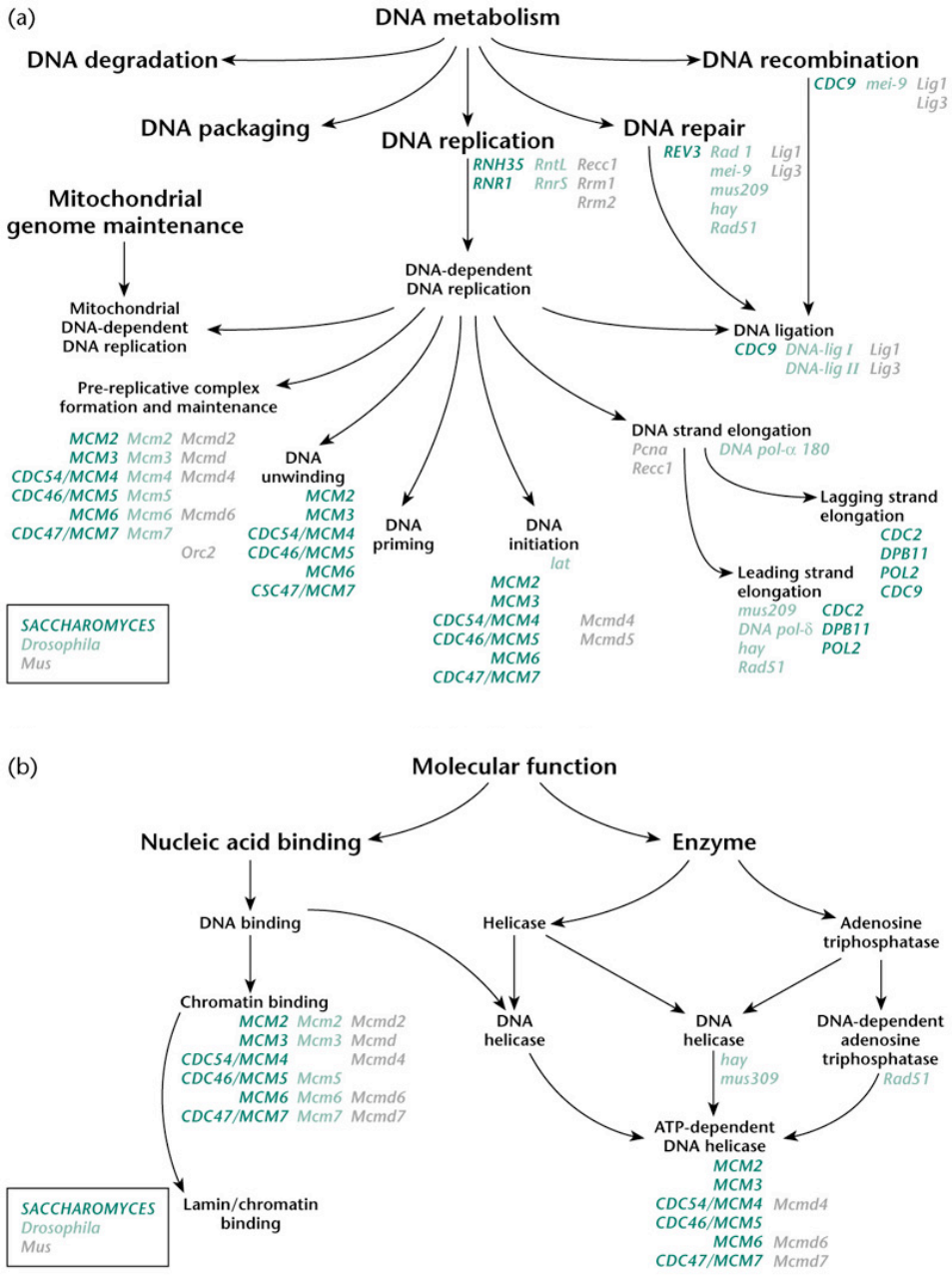
Cellular Component

- Host cell nucleus (GO:0042025)

Identification of gene function (contd)

- You have identified a gene – what is its function? (contd)
  - Gene ontology (GO) – highly structured vocabulary for gene classification
    - Genes are classified using this vocabulary
    - Relates protein function with cellular or organismal functions
      - Nucleic acid binding
      - Cell division

- HUGE CAVEAT!
  - Genome must be well annotated or else GO terms are absent or incomplete



Genome annotation

- Extremely important as number of sequences increases
  - Goals are to identify
    - all of the sequences
    - all of the features of each sequence
    - All of the functions of the identified genes
  - Sometimes annotation does not agree with known function
    - Human error
    - New and updated information not propagated to database
    - Inaccurate sequencing
    - Sometimes annotation is correct but protein lacks function under certain conditions (e.g., need cofactors)
  - Gold standard for functional analysis is loss-of-function analysis
    - Most accurate annotation
  - Common to have "annotation jamborees" where biologists and bioinformaticians come together to annotate new sequences
    - *Xenopus tropicalis* jamboree was in Spring 2006
    - But many genes and gene models are still unannotated

