BioSci D145 Lecture #5

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

- TA – Riann Egusquiza (regusqui@uci.edu)
  - 4351 Nat Sci 2– office hours M 1:45-3:45
  - Phone 824-6873

- check e-mail daily for announcements, etc

- Updated lectures will be posted on web pages after lecture
  - http://blumberg-lab.bio.uci.edu/biod145-w2018
  - http://blumberg.bio.uci.edu/biod145-w2018/

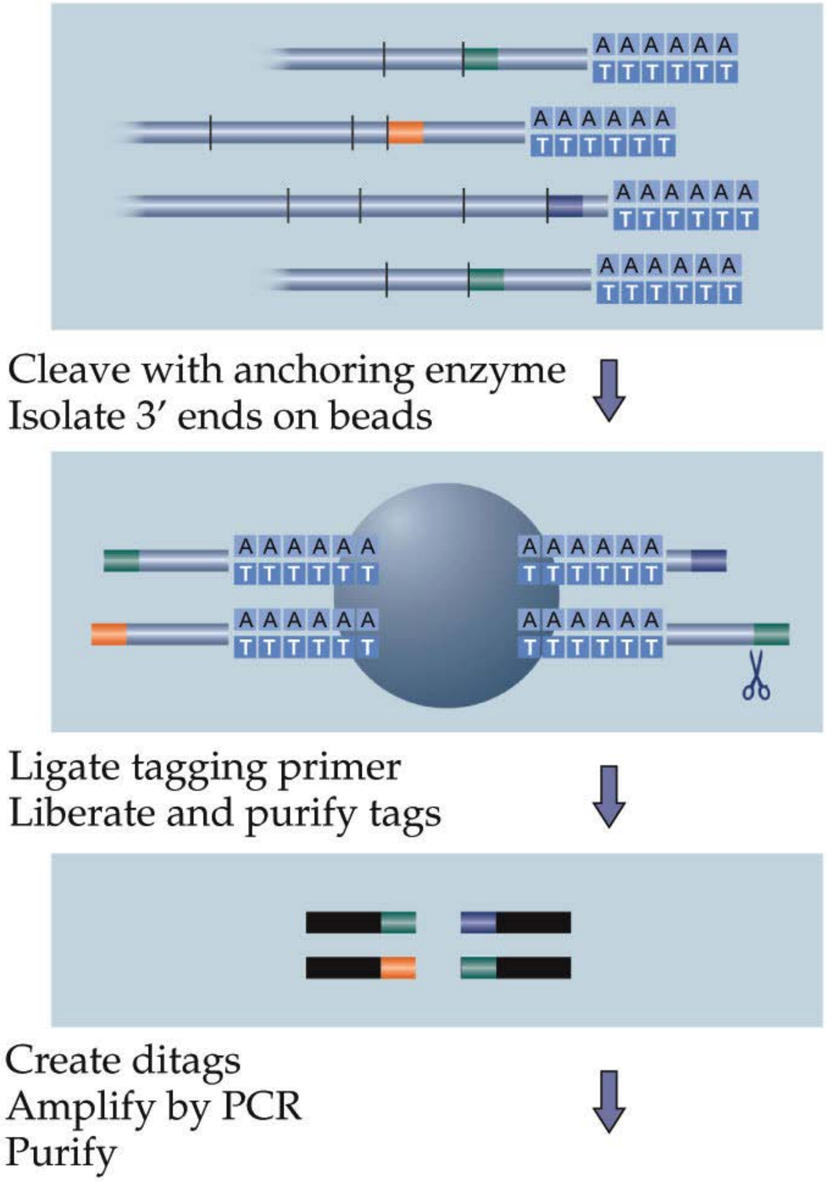
- Last year's midterm is posted.

- Answers to last year's midterm will be discussed at end of today's class, or posted if we don't get there.

Other methods of transcriptome analysis - parallel

- Microarray was once the dominant method
  - Sequencing-based methods have almost totally replaced microarrays
  - SAGE (serial analysis of gene expression)
    - Nanostring is modern implementation
    - Short sequences
  - RNAseq
    - Directly sequence large numbers of RNAs
    - Longer sequences

- SAGE
  - Relies on generating many very short sequences and matching these to the genome
  - 10 bp = short SAGE
  - 17 bp = “long” SAGE



Cleave with anchoring enzyme
Isolate 3′ ends on beads

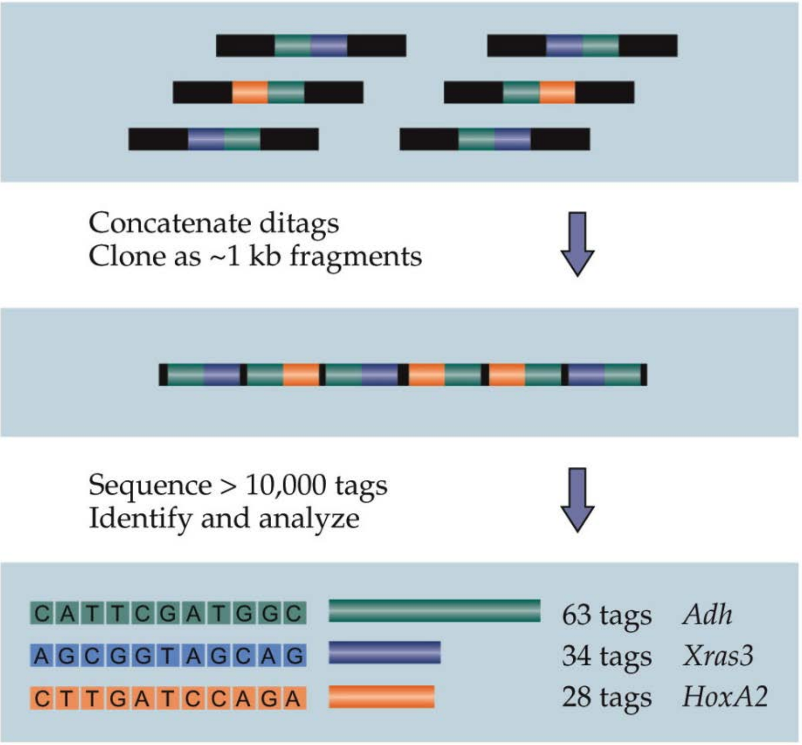
Ligate tagging primer
Liberate and purify tags

Create ditags
Amplify by PCR
Purify

SCIENCE 3e, Figure 4.14 (Part 1)

Other methods of transcriptome analysis - parallel

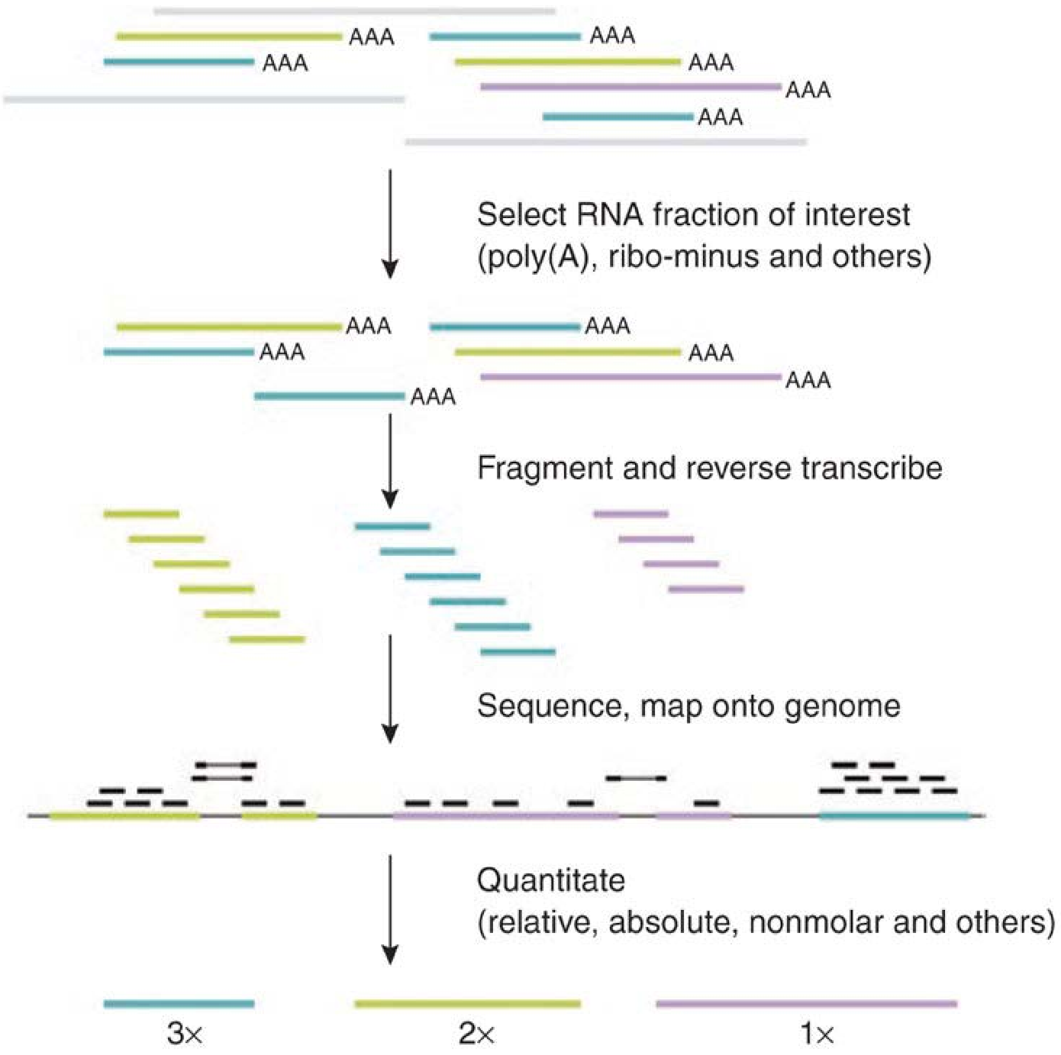
- SAGE (continued)
  - What is the obvious shortcoming of this method?
  - Sequences may not be unique and could have difficulty mapping to the genome



GENOME SCIENCE 3e, Figure 4.14 (Part 2)

Other methods of transcriptome analysis - parallel

- RNA seq – Ali Mortazavi is local expert
  - Use of massively parallel sequencing allows precise quantitation of transcript
  - Also allows discovery of rare splice forms
  - Discovery of unexpected transcripts
  - Main problem is in mapping sequence calls to genome
    - Sequencing has 1-2% errors which can make mapping to genome fail
    - or induce “in silico cross-hybridization”
      - Mapping to incorrect genomic location


AAA
Select RNA fraction of interest
(poly(A), ribo-minus and others)
Fragment and reverse transcribe
Sequence, map onto genome
Quantitate
(relative, absolute, nonmolar and others)
3×
2×
1×

Microarray vs. RNAseq

- Microarray
  - Assumes you know all the transcripts
  - Any sequence you did not know was expressed will not be there.
    - except whole genome tiling arrays – Kapranov paper
  - Detection limit issues
    - Signal-noise ratio
  - Well validated , expression analysis can be quantitative

- RNAseq
  - No assumption re transcripts but best to have genome sequence (can do de novo assembly)
  - Can discover novel sequences or new splice forms not yet characterized (if you have genome)
  - Detection limits are not a problem – can detect small #
  - Getting better, expression analysis can be quantitative

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)
    - Microarrays (Oligo, cDNA, protein)
  - Novel genes
    - Differential display
    - Expression profiling
      - SAGE and related approaches

- What do they interact with?
  - Biochemical methods
  - Yeast two, three hybrid screening
  - Phage display
  - Expression cloning
  - Proteomics
    - 2 dimensional gel electrophoresis
    - Mass spectrometry
    - Protein microarrays

Methods of profiling gene expression (small number of genes)

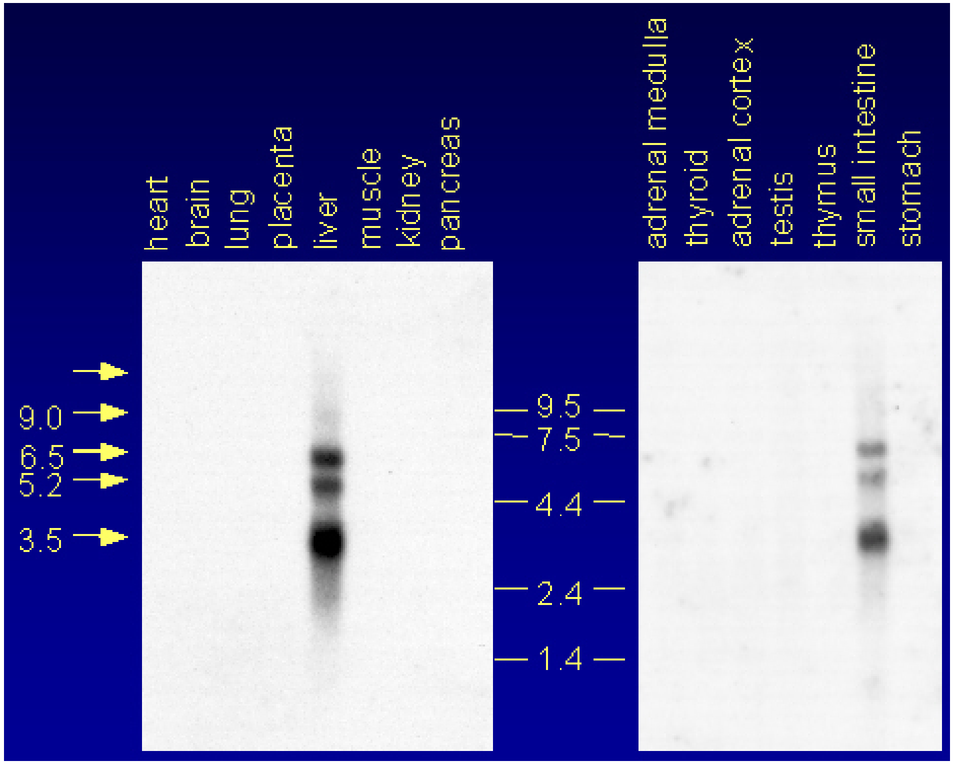
- How to evaluate gene expression?
  - Old, low-throughput - prepare RNA sample and perform
    - Northern blot – immobilize RNA on filter, probe
      - Quantitative WHY?

Probe is in excess

    - Nuclease protection
      - quantitative
    - In situ hybridization
      - Not quantitative – enzymatic reaction
  - Newer, high throughput methods
    - RT-PCR
      - Can be quantitative
    - Quantitative real time RT-PCR
  - Or prepare protein samples and evaluate proteins
    - Western blot - detect protein of interest with specific antibody.
    - ELISA – enzyme linked immunosorbent assay quantitative
    - RIA – radioimmunoassay - quantitative

Analysis of mRNA - size and splicing

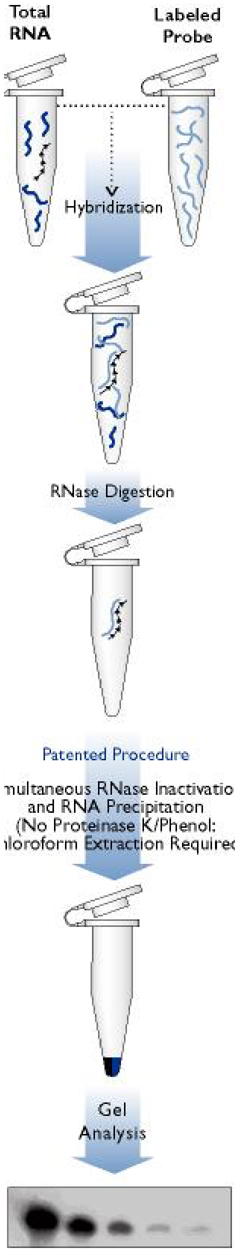
- Quantitation of mRNA levels
  - possible methods
    - Northern analysis
    - nuclease protection
    - RT-PCR
  - measure steady state mRNA levels (production/degradation)



- mRNA size determination –
  - Northern blot only way
  - good RNA size markers = accurate sizing
  - which to use, poly A^+ or total RNA?
    - A^+ much more sensitive (50-100x)
      - what about mRNAs with no or short tails?
    - total RNA much simpler
      - gel limitations – 20 µg/lane is practical limit
  - what is a key factor in sizing mRNAs?

Appropriate size standards larger and smaller than target

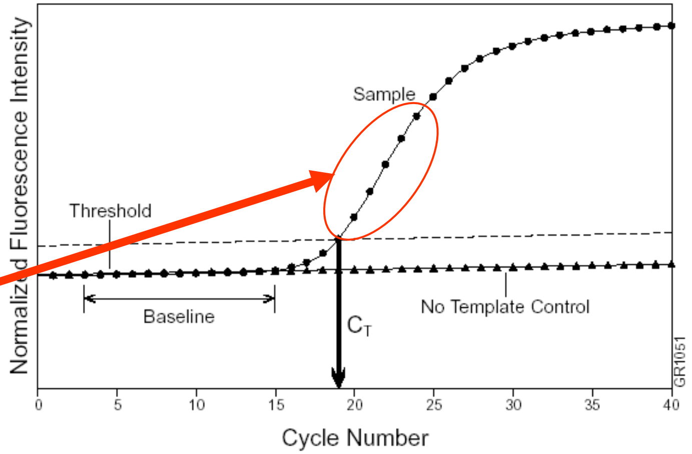
Analysis of mRNA - quantitation (contd)



- Nuclease protection assays
  - approach
    - hybridize a single-stranded (SS) probe (DNA or RNA) to RNA sample
      - probe must be larger than protected region
    - digest remaining single stranded regions
    - electrophorese on denaturing polyacrylamide gel
  - advantages
    - less sensitive to slightly degraded mRNA
    - absolutely quantitative
    - can tolerate large amounts of RNA (100+ µg)
      - allows detection of rare transcripts
      - but gives high background
    - multiple simultaneous detection
  - disadvantages
    - more tedious than Northern
    - no blot to reuse
    - multiple simultaneous detection hard to optimize

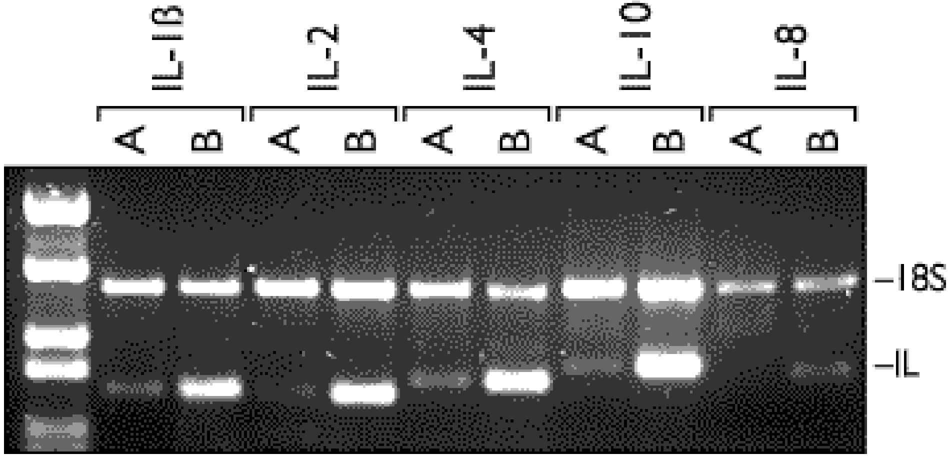
Analysis of mRNA - quantitation (contd)

- RT-PCR - reverse transcriptase mediated PCR
  - approach
    - reverse transcribe mRNA -> cDNA
    - amplify with specific primers
    - quantitate
  - flavors
    - relative quantitation – compare to invariant gene
    - absolute quantitation
      - by comparison to synthetic reference
      - competitive PCR
      - various fluorescent dye mediated methods
  - advantages
    - very fast and simple
    - works with tiny amounts of material
  - limitations
    - RT efficiency differs by mRNAs
    - Must be in linear amplification range
    - Errors increase exponentially with amplification



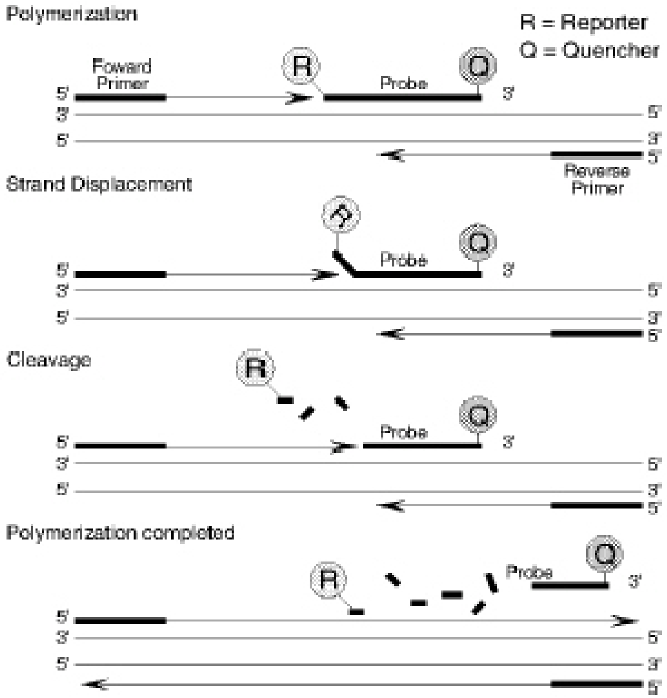
Analysis of mRNA - quantitation (contd)

- RT-PCR reverse transcriptase mediated PCR
  - relative concentration determination
    - perform multiplex reaction using two primer sets
      - 1 for reference, 1 experimental
    - advantages
      - no fancy equipment required
    - disadvantages
      - careful attention to linear region for both primer sets
      - often must add one set during reaction
        - » companies claim to have products that eliminate this need
        - » more than 2 primer sets are not reliable



Analysis of mRNA - quantitation (contd)

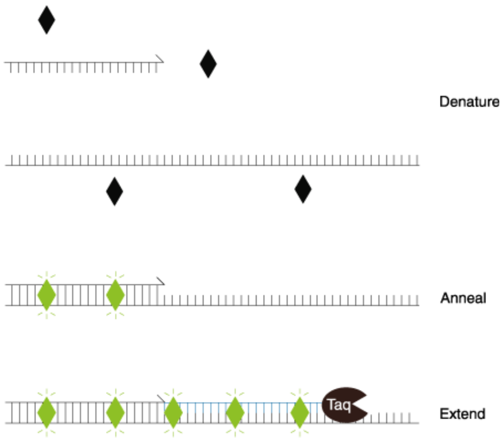
- RT-PCR (contd)
  - absolute concentration determination real time PCR
    - Taqman, molecular beacons
      - Fluorescent methods that allow direct quantitation of PCR product
    - approach
      - special oligonucleotide that has a fluor and a quenching group on it.
        - » When whole, no fluorescence
      - perform PCR reaction, if primer anneals, Taq polymerase removes the reporter group which can now fluoresce



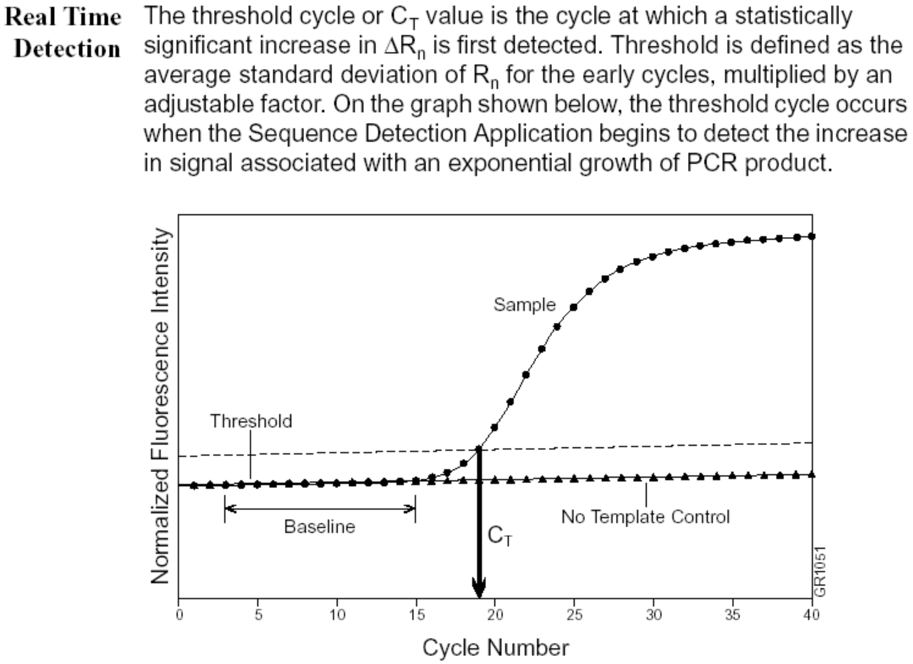
Analysis of mRNA - quantitation (contd)

- RT-PCR (contd)
  - absolute concentration determination - Taqman, etc
    - Fluorescence detected continuously in real time
    - advantages
      - can be detected in real time with proper instrument
      - no difficulties with linearity
      - multiplexing of probes possible (limited by available dyes)
      - very good for clinical diagnostics
    - disadvantages
      - requires instrument
        - » varies from expensive to extremely expensive
        - » Not of equal quality
      - need to make custom oligos - can be expensive
      - must know something about relative abundance of mRNAs before setting up reactions
      - careful optimization required for best results
        - » primer concentrations
        - » target concentrations

Analysis of mRNA - quantitation (contd)



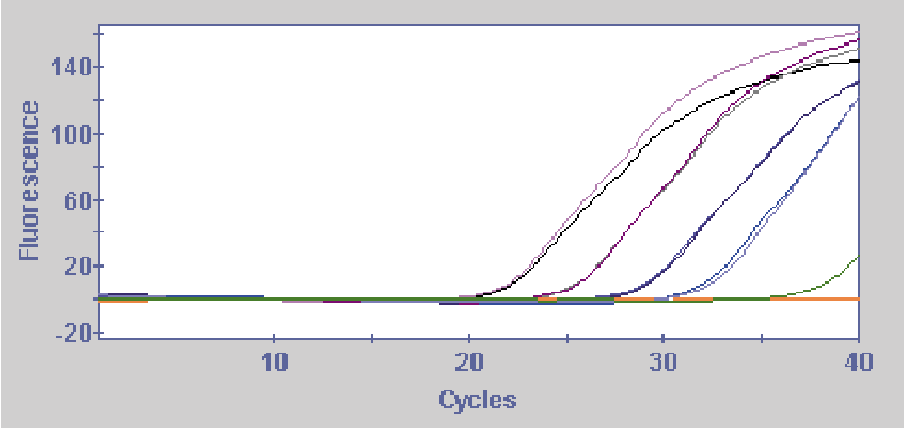
- RT-PCR (contd)
  - absolute concentration determination – Sybr Green
    - Alternative real time RT-PCR utilizes a single dye
    - approach
      - Extend a single template
      - Detect ds DNA with a specific dye



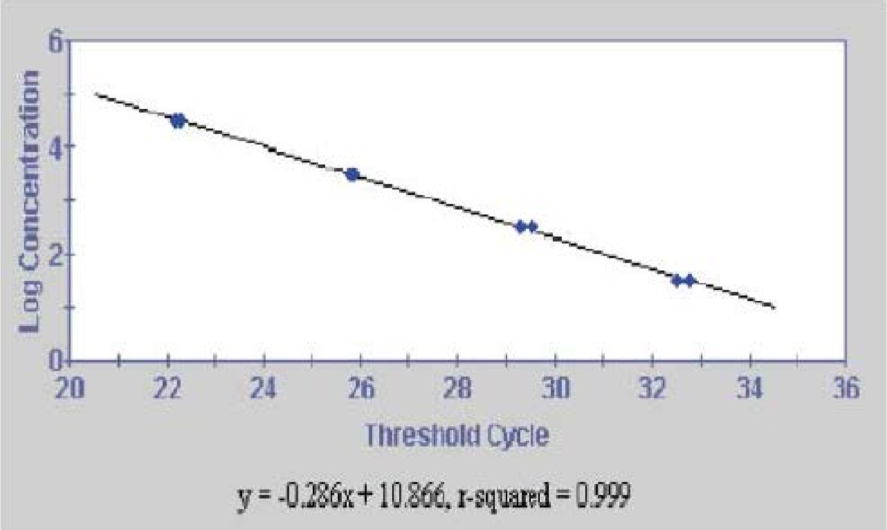
Real Time Detection

The threshold cycle or C_T value is the cycle at which a statistically significant increase in ΔR_n is first detected. Threshold is defined as the average standard deviation of R_n for the early cycles, multiplied by an adjustable factor. On the graph shown below, the threshold cycle occurs when the Sequence Detection Application begins to detect the increase in signal associated with an exponential growth of PCR product.

Analysis of mRNA - quantitation (contd)

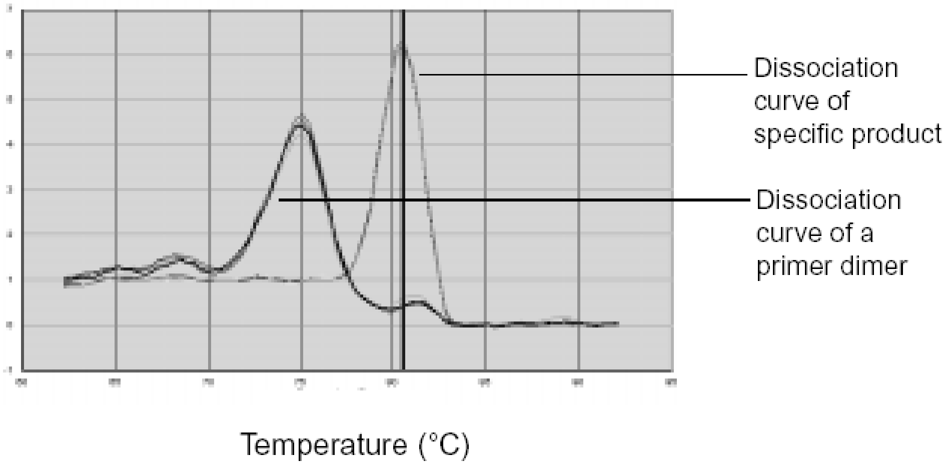


- RT-PCR (contd)
  - absolute concentration determination – Sybr green
    - Plot lift off time
    - Generate standard curve



Analysis of mRNA - quantitation (contd)

- RT-PCR Sybr Green (contd)
  - Advantages
    - No special primers needed
    - Single dye, simple
    - Fast, robust and quantitative
    - Good for routine use
  - Disadvantages
    - Need instrument
    - Single dye, can’t multiplex
    - Problems with multiple fragments
      - » Melting curves required
    - Absolute quantitation requires std curve



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 copy of genes derived from duplication or mating event, e.g., duplicated genes in tetraploid organisms

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

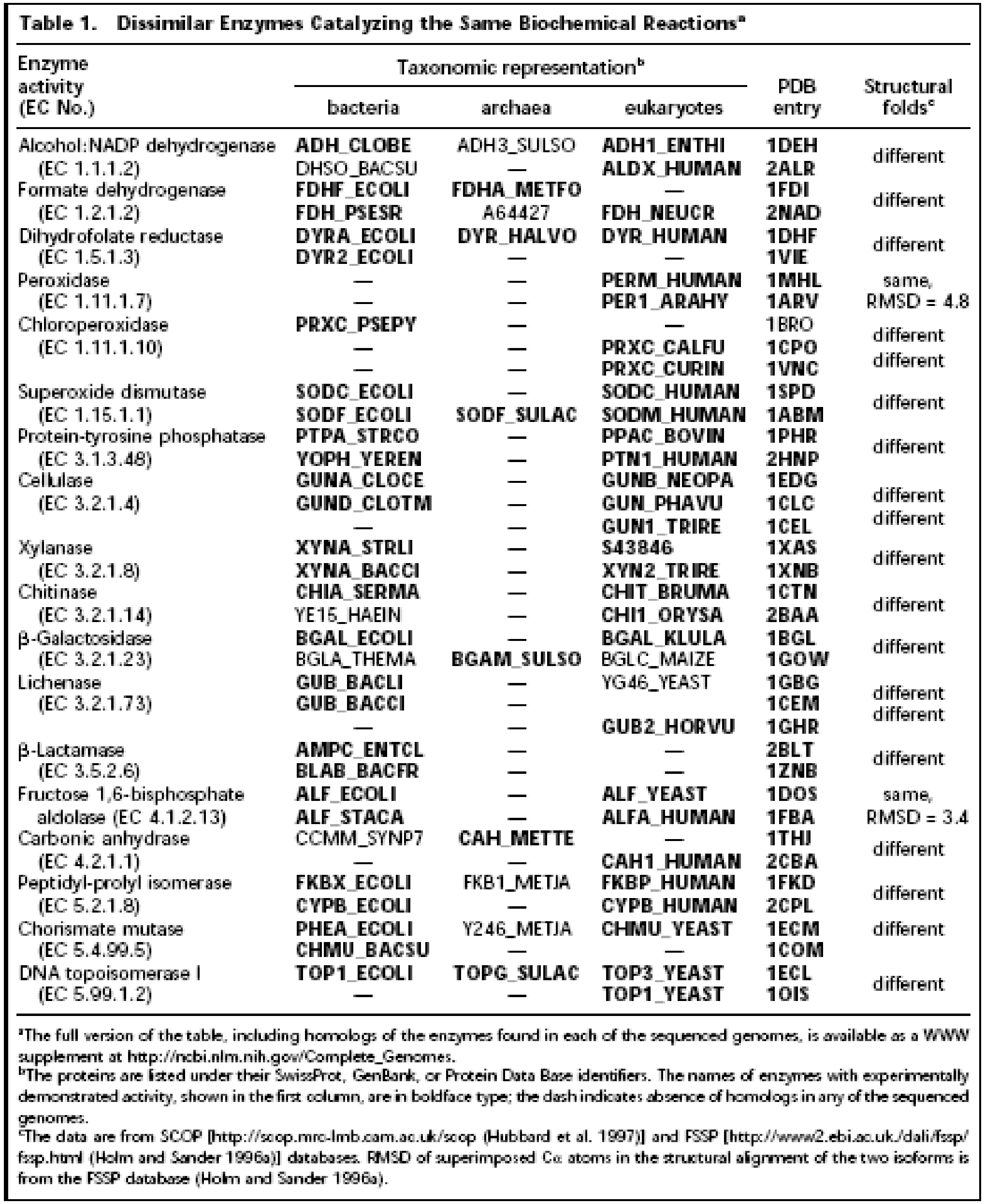


Table 1. Dissimilar Enzymes Catalyzing the Same Biochemical Reactions[a]

Enzyme activity (EC No.)	Taxonomic representation[b]			PDB entry	Structural folds[c]
	bacteria	archaea	eukaryotes		
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, RMSD = 4.8
(EC 1.11.1.7)	—	—	**PER1_ARAHY**	**1ARV**	
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, RMSD = 3.4
aldolase (EC 4.1.2.13)	**ALF_STACA**	—	**ALFA_HUMAN**	**1FBA**	
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)

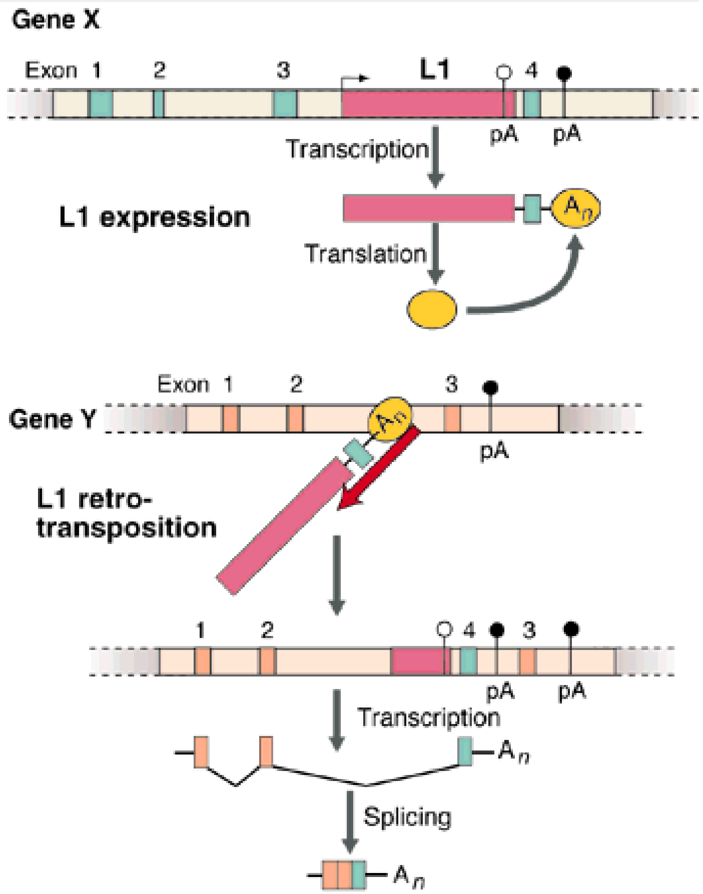
- Genes with very different functions can be related
  - 3-D structure may indicate that proteins are related (evolved from the same ancestral protein) but sequence identity too low to detect
    - Expected when genes diverge from a distant common ancestor
    - < 20% amino acid sequence identity too 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

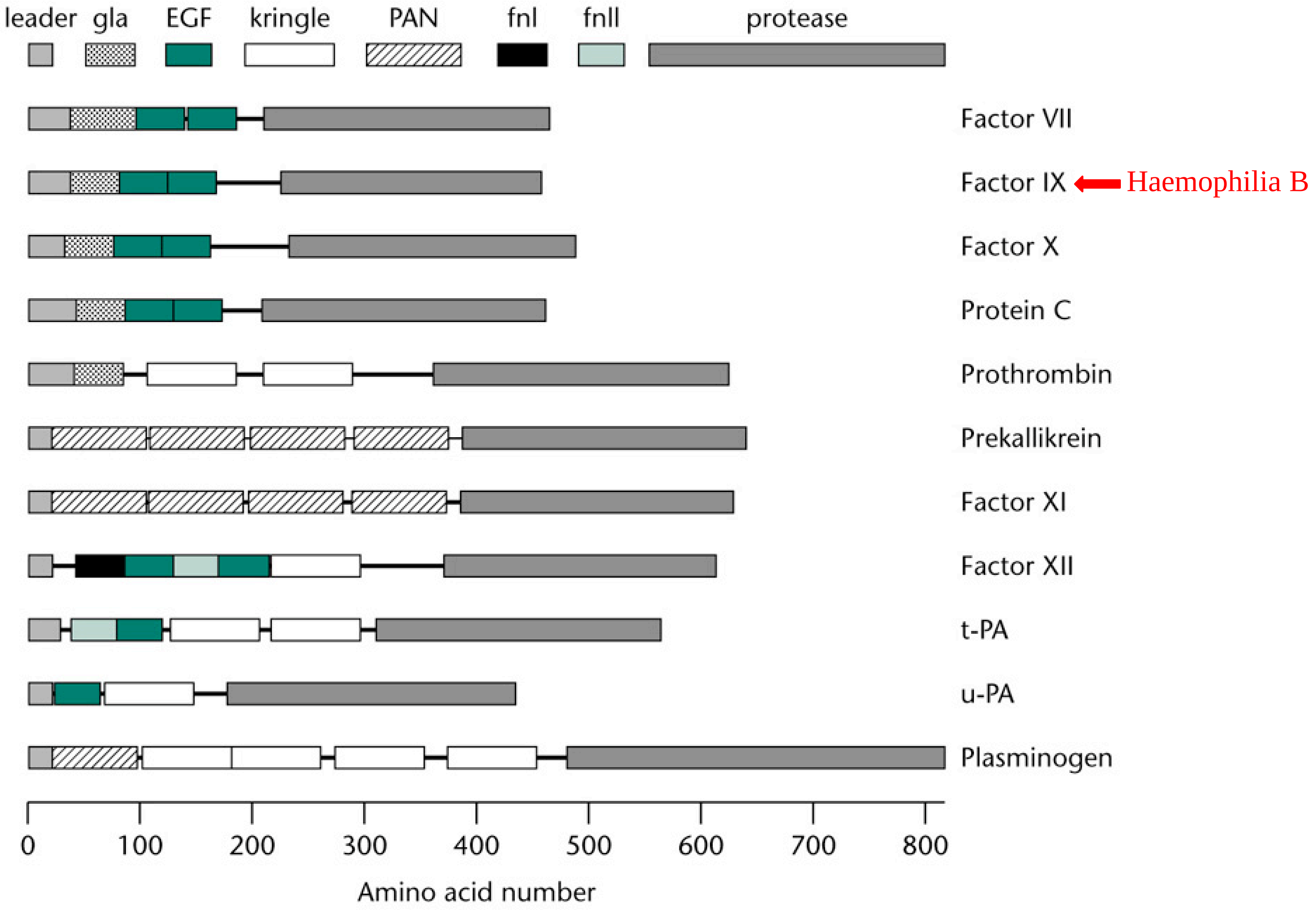
Comparative genomics (contd)

- 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 antiobiotic 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)
    - Microarrays (Oligo, cDNA, protein) (Iyer)
  - Novel genes
    - Expression profiling
      - Genomic tiling microarrays (Kapranov)
      - SAGE and related approaches (RIKEN)
      - Massively parallel sequencing (RNA-Seq) (Bentley)
- Which genes regulate what other genes? (week 6 papers)
- Epigenetic modification of gene expression (week 7 papers)
- What is the phenotype of loss-of-function? (week 8 papers)
  - Genome wide CRISPRi (Liu)
  - Genome wide synthetic lethal screens (Luo)
  - CRISPR/Cas (Gilbert)
- What do they interact with (week 9 papers)
- Metabolome & microbiome (week 10 papers)

1. **(8 points)** Did you know that there are carnivorous plants that survive in nutrient poor environments by eating insects? Among these are three types of "pitcher plants", that all trap insects by drowning them in a sweet liquid contained in a modified leaf that looks like a pitcher. Interestingly, the Australian, Asian and American pitcher plants all look very similar and catch insects the same way. However, they are believed to be completely unrelated biologically. The Australian pitcher plant is thought to be related to star fruit, the Asian pitcher plant to buckwheat and the American pitcher plant to kiwifruit. Your group's mission is to determine 1) whether this is an example of convergent evolution or whether the plants are similar but have been misclassified and 2) what types of adaptations allow these plants to digest insects to extract nutrients such as phosphorous and nitrogen.

   a) (4 points) **What approach would you take to determine whether these pitcher plants are closely related to each other or not? How will you place them among the evolutionary tree of plants and confirm or refute the classification of taxonomists?**

Since you want to determine how closely related these plants are, and specifically study their functional adaptations (in b), the best answer would be to perform whole genome sequencing for the 3 types of pitcher plants. It is 2017, so you will want to perform Nextgen sequencing, most likely by Illumina Solexa sequencing. Isolate DNA, generate Illumina libraries, sequence each genome to high depth of coverage and assemble them with standard bioinformatic tools to generate draft genome sequences. Then compare these sequences with each other to determine how closely related they are and then with sequences known from other plants to accurately place these pitcher plants on the plant phylogenetic tree. Check whether your classification matches that of taxonomists.

b) (4 points) One hypothesis is that the plants harbor specific microorganisms in their "pitchers" that enable them to extract nutrients from the insects, not dissimilar from gut bacteria that enable primates to digest fiber to produce short-chain fatty acids. An alternative hypothesis holds that the plants have modified proteins that were originally responsible for cellular defense to produce digestive enzymes that break down insects. **What approach could you take to 1) determine whether the microbial contents differ significantly between pitcher plants in the same species and among the 3 different types of pitcher plants? How could you test the hypothesis that a common cellular enzyme such as purple acid pyrophosphatase has specific amino acid changes in carnivorous, vs. related non-carnivorous plants?**

To address whether the microbiomes differ among plants, collect samples from several (~5) individuals of each species, isolate DNA and perform environmental shotgun sequencing, much like the Venter paper (but use Nextgen sequencing). Compare the sequences in each species and between species to identify any potential similarities and differences.

To test the hypothesis that specific changes in common enzymes are found in carnivorous, vs. non-carnivorous plants, simply compare the sequences between related carnivorous and non-carnivorous plants and with other plants. This was actually done and showed that there were common substitutions in totally different lineages that facilitated a carnivorous mode of obtaining nutrients.

2. (4 points) In an even more bizarre evolutionary development, the Asian pitcher plant *Nepenthes hemsleyana* has abandoned catching insects for food and instead has developed a mutualistic relationship with the wooly bat. *N. hemsleyana* doesn't produce much fluid in its pitcher and has developed a shape perfectly complementary to that of the bat such that the bats roost inside the plant. The bats defecate inside the plant, providing the plant with nutrients. The closely related species, *N. raffiesiana* lives in the same environment and catches insects in the usual way to obtain nutrients. **Please describe how would you identify potential gene candidates that enable *N. hemsleyana* to attract bats and utilize their feces for nutrition compared with *N. raffiesiana*?**

Since you have two closely related species (and already sequenced one of them) it would be relatively simple to sequence the other and compare what differences are found between them. This will identify candidate genes that you could use for future studies, perhaps after selecting those known to be related to nitrogen and phosphorous metabolism and uptake. The key point is to sequence both and do a detailed comparison.

3. (8 points) There is a genus of lizards, *Geckolepsis*, commonly referred to as the fish scale geckos which are found only in Madagascar. This week, a paper was published describing a new species, *Geckolepsis megalepsis*, that has gigantic scales that can rapidly detach when the lizard is attacked by a predator. The predator is left with a mouthful of scales while the lizard gets away and regenerates its skin and scales perfectly (i.e., without scarring) in a few weeks. Other species in the *Geckolepsis* genus have large scales (although not as large as *G. megalepsis*) but lack this rapid detach/regenerate mechanism - they can lose a few scales but regenerate them imperfectly. Your group's mission is to identify how *G. megalepsis* can detach and regenerate its skin and scales while the spotted fish scale gecko, *G. maculata* cannot.

a) (4 points) An obvious starting point would be to sequence the genomes of *G. megalepsis* and *G. maculata*. One of the TAs, Ron, has given you an Applied Biosystems 377 capillary sequencer and 4 PCR machines and suggests that you use these to do cycle sequencing of the genomes as pioneered by Craig Venter in his Sargasso Sea paper that we read. **Is Ron correct? Can this approach generate complete genome sequences in one quarter? If he is correct, please explain why. If he is not correct, please describe succinctly how you will produce a high quality draft sequence in one quarter.**

Ron is incorrect. A capillary sequencer is for Sanger sequencing, not Nextgen sequencing so you will not be able to come close to even a fragment of one genome in a quarter – it simply does not have enough capacity for rapid, whole genome sequencing. I will isolate DNA from the two species of interest, fragment them up to make Illumina sequencing libraries and do enough sequencing runs to generate the entire sequencing. This could be as few as a single run, depending on the instrument available. Let the computer assemble this sequence and produce draft genomes. If you are very industrious, your group might consider adding a different sequencing method (such as 454 or PacBio) to help resolve gaps.

b) (4 points) The approach your group took in a) was partially successful - you generated draft genome sequences but these are highly fragmented. The estimated total genome size is 1.4 gigabases, about half of human. There are 24 chromosomes, but your analysis generated more than 10,000 scaffolds for each species. Oops. Ron suggests that you quickly generate a radiation hybrid map of the two genomes to facilitate the assembly since the large phenotypic difference between two closely related species suggests that there may only be a small number of actual changes. **Is Ron correct? If so, please say why and what you will need to generate a good RH map. If he is not correct, please explain why and what method you would use to generate a high quality genome map that will allow you to assemble the genome. In either case, what markers will you use and how will you obtain them?**

Once again, Ron is incorrect (why is he your TA anyway?) He is wrong because it is not possible to quickly generate a radiation hybrid panel and map – this could easily take years. I would instead generate a BAC library from each species, use these for BAC end sequencing (using old fashioned Sanger sequencing) and then use these STCs as markers. Generate the map by comparing the BAC end sequences with your draft genome to see which pieces go where. This will probably take longer than one quarter, though. Alternatively, you could generate unique markers from your genome sequencing and perform HAPPY mapping. This would be less accurate, but perhaps a bit quicker. Either answer is ok if you described how it could achieve your goals and what markers you used.

4. (15 points) *Geckolepsis* are classified with the largest subgroup of the Family *Gekkonidae*. *Gekkonidae* are found worldwide, but are particularly diverse and species-rich in tropical areas. Since the diversity of this group is so large, you might reasonably infer that the rate of evolution within the *Gekkonidae* is unusually high.

a) (5 points) The next task is to generate a very precise phylogenetic analysis of representative member of the Family *Gekkonidae* and the entire Genus *Geckolepsis* (which has 5 species). Ron suggests that a microarray analysis would be the most accurate and fastest way to generate an accurate phylogenetic tree. The other TA, Riann says that Ron is wrong, but doesn't tell you why. **Is Ron correct or not? If he is correct, outline how you will perform the phylogenetic analysis and determine which lizards have conserved, constrained or rapidly evolving regions of their genomes. If Riann is correct, state why and then outline how you would perform the same analysis.**

Ron is still incorrect while Riann is correct. Although there are such things as "phylogenetic microarrays" we did not discuss them and there is no possibility that they will be the most accurate and fastest way to generate an accurate phylogenetic tree for the Gekolepsis genus together with representative members of the Gekkonidae. The best approach would be like what was done in the Lindblad-Toh paper. Collect the available reptile genomic sequences, then identify which species you will choose from other groups as well as Gekkonidae groups and the 5 species of Geckolepsis. Generate draft genomes of these, build phylogenetic trees by computer and analyze to identify conserved, constrained and rapidly evolving regions of the genome.

b) (5 points) Ron is getting pretty bossy (for a TA) and next decides that you should look for copy number variations in the genomes of the 5 species of Geckolepsis. Your group doesn't want to do any extra work and debates whether you should listen to Ron, or instead start ignoring him and talk to Riann instead**.** **Will the analysis you have done in 4a be able to reveal most or all of the copy number variations in the 5 species? If so, please explain why and what aspects of the analysis you did in a) will provide this information so that you can move on to the next task. If it will not, please say why not and how you would go about identifying most or all of the copy number variations in the 5 species. Be sure to say what materials you needed for your analysis.**

It is unlikely that the genome sequences will reveal copy number variations, although, they could give some idea about whether such variations exist. You will want to generate genome tiling microarrays like in the Redon paper and use these to identify all of the copy number variations in the 5 species of Geckolepsis. With such microarrays, you will also be able to identify CNVs among individuals within a species.

c) (5 points) Unfortunately, neither the phylogenetic analysis in a), nor the CNV analysis in b) identified why and how G. megalepsis is able to shed its scales/skin and easily escape predators AND regenerate both skin and scales perfectly. Clearly the next step is to ask whether the profile of RNA transcripts differs in the skin of G. megalepsis vs. G. maculata. Once again, Riann and Ron are offering conflicting advice - Ron wants you to use microarray analysis and Riann says that RNA-seq is the way to go. 4 people in your group vote for microarrays and 4 for RNA-seq - you have to break the tie. Both methods will work to some extent - your TAs are smart people after all. **Your goal is to identify all of the transcripts responsible for the ability of G. megalepsis to shed/regenerate its skin and scales. Please describe how you will tackle this problem and why your approach will provide the best chance to identify the gene(s) responsible.**

Since you want to identify ALL of the transcripts that could be responsible for the ability of G. megalepsis to shed/regenerate its skin and scales, RNA-seq is really the only choice. This is because any sort of expression microarray will use only expressed genes as targets on the chip. In contrast, RNA-seq can identify any sort of transcript, irrespective of whether it is mRNA, lcRNA, microRNA or any other sort of strange RNA. I would prepare RNA from the skin of G. megalepsis and G. maculata before and after injury, then perform RNA-seq and compare which genes are expressed before and after injury. An acceptable, although probably less effective, approach would be to use whole genome tiling arrays to analyze which transcripts are produced and perform the same sort of comparison.