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 1-3
  - Phone 824-6873

- check e-mail and noteboard daily for announcements, etc..
  - Please use the course noteboard for discussions of the material

- Updated lectures will be posted on web pages after lecture
  - http://blumberg-lab.bio.uci.edu/biod145-w2017

- 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.

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

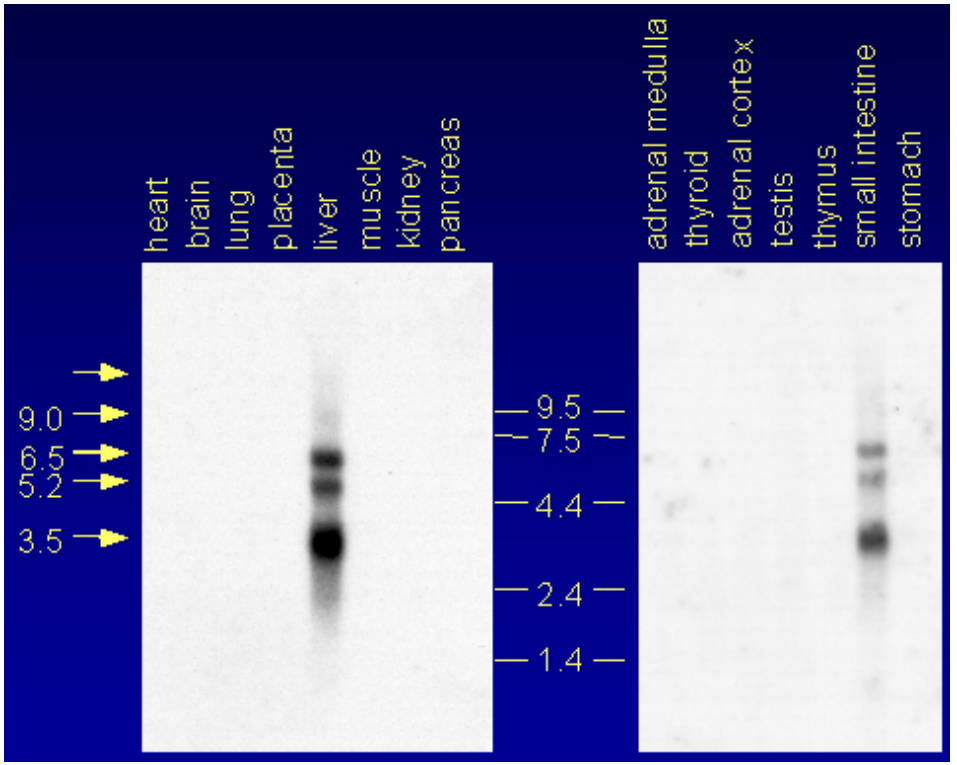
Methods of profiling gene expression (small number of genes)

- 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?

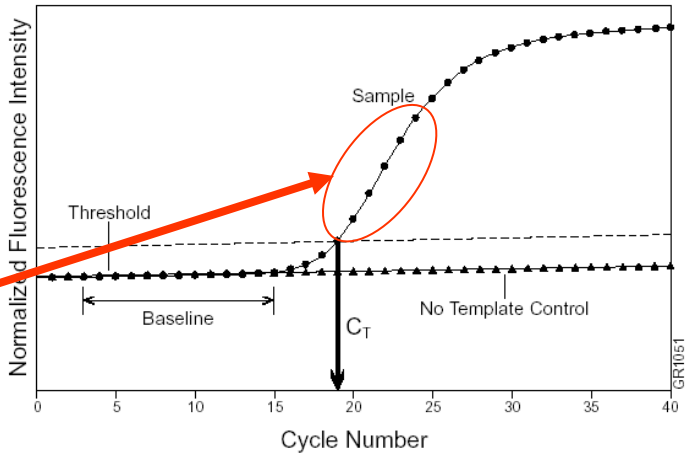
Appropriate size standards larger and smaller than target

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

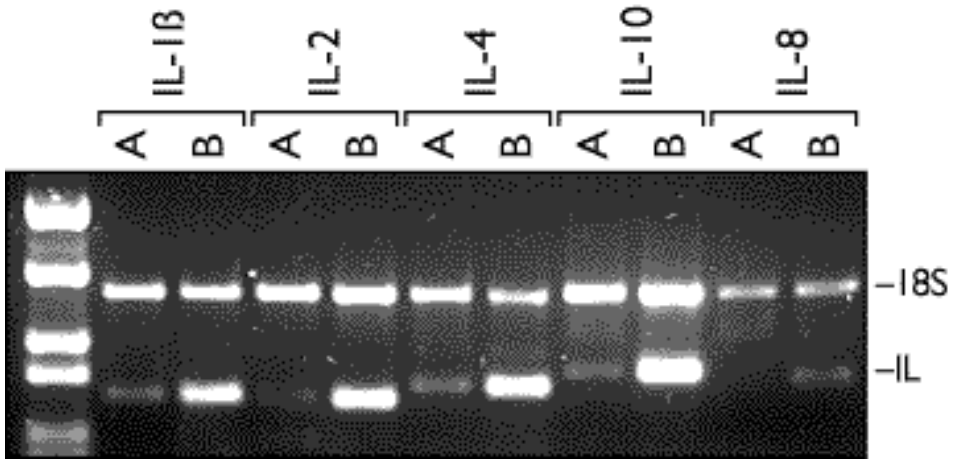
Analysis of mRNA - quantitation (contd)

- 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



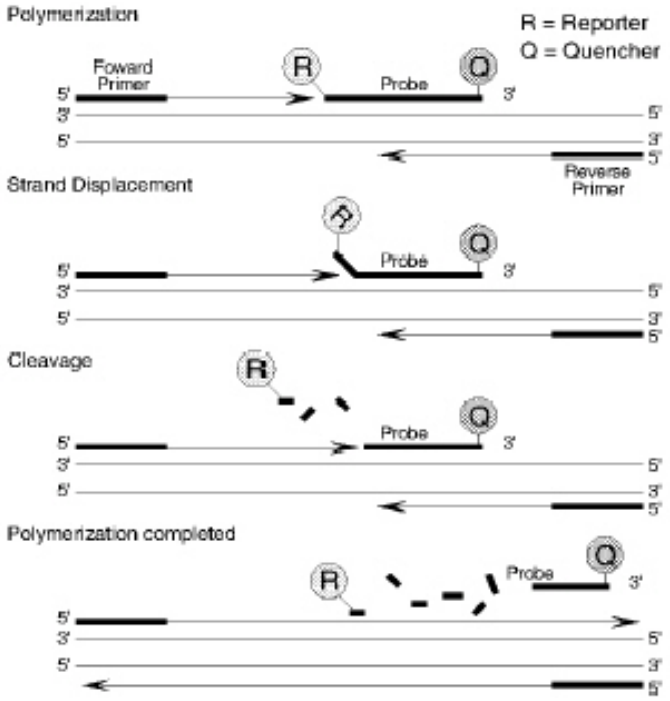
Analysis of mRNA - quantitation (contd)

- 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)

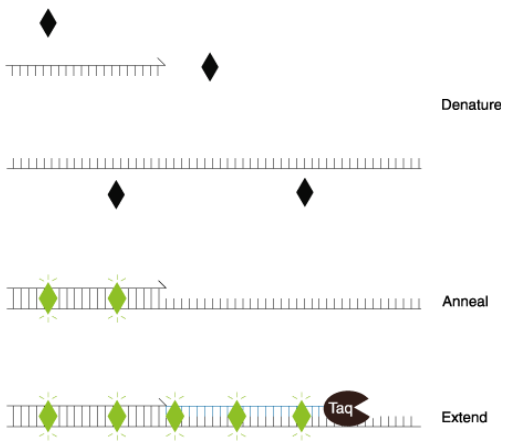
- RT-PCR (contd)
  - absolute concentration determination real time PCR
    - Taqman, molecular beacons
      - Fluorescent methods that allow direct quantitation of PCR product
    - approach
      - special oligonucleotide that has a fluor and a quenching group on it.
        - » When whole, no fluorescence
      - 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

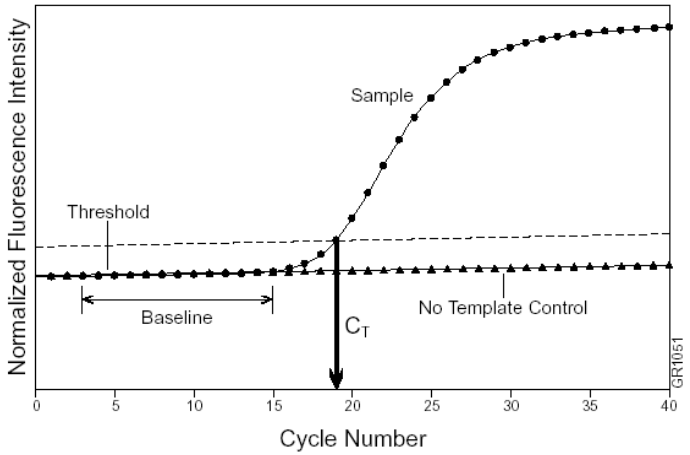
Analysis of mRNA - quantitation (contd)



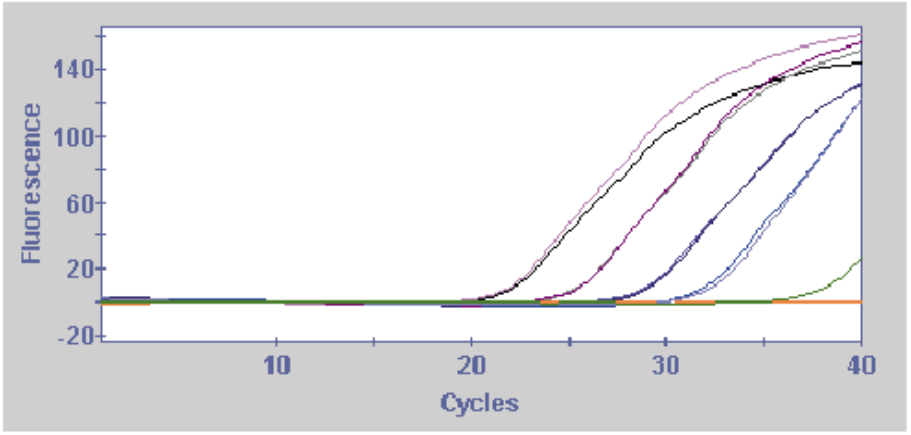
- RT-PCR (contd)
  - absolute concentration determination – Sybr Green
    - Alternative real time RT-PCR utilizes a single dye
    - approach
      - Extend a single template
      - Detect ds DNA with a specific dye

Real Time Detection

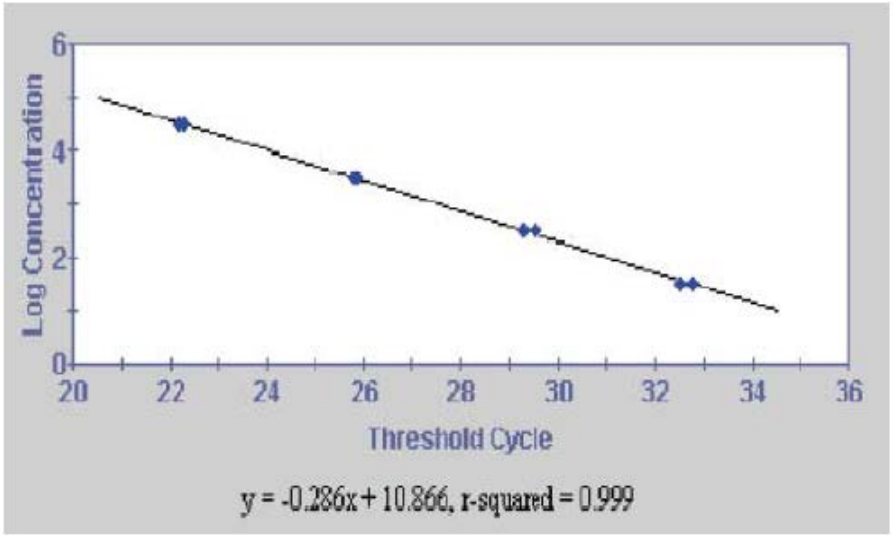
The threshold cycle or C_T value is the cycle at which a statistically significant increase in ΔR_n is first detected. Threshold is defined as the average standard deviation of R_n for the early cycles, multiplied by an adjustable factor. On the graph shown below, the threshold cycle occurs when the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



Analysis of mRNA - quantitation (contd)

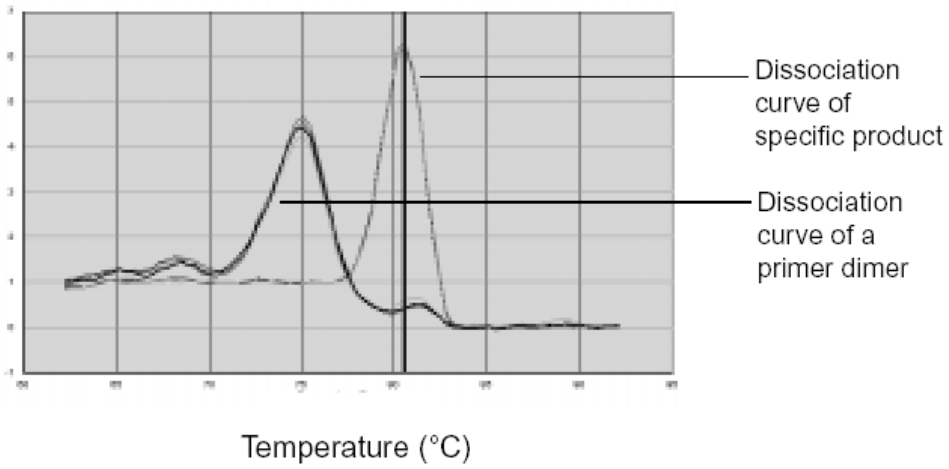


- RT-PCR (contd)
  - absolute concentration determination – Sybr green
    - Plot lift off time
    - Generate standard curve



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] 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)

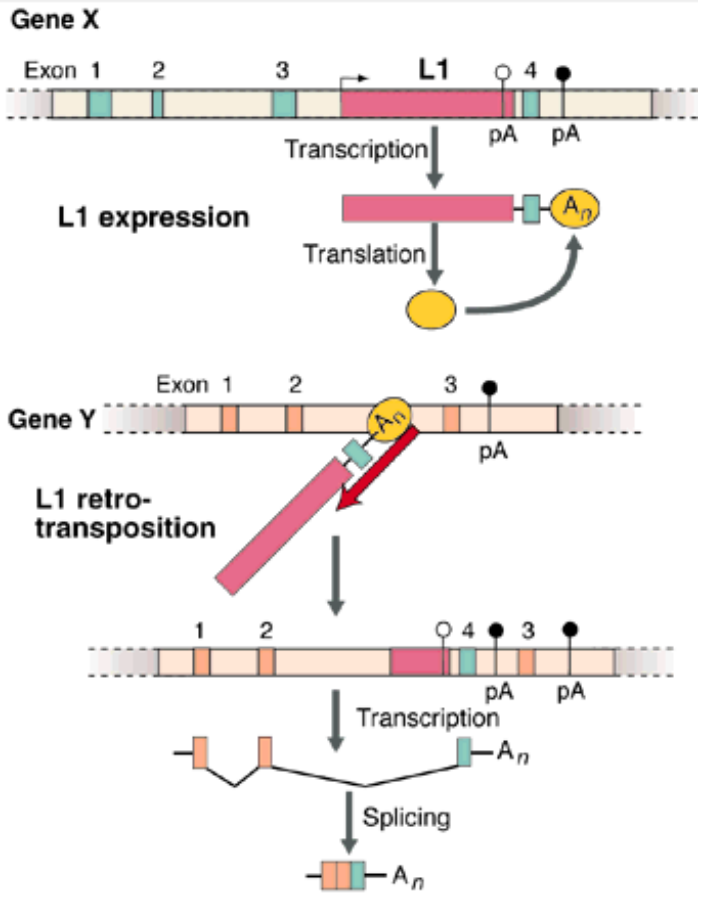
- Genes with very different functions can be related
  - 3-D structure may indicate that proteins are related (evolved from the same ancestral protein) but sequence identity too low to detect
    - Expected when genes diverge from a distant common ancestor
    - < 20% amino acid sequence identity too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

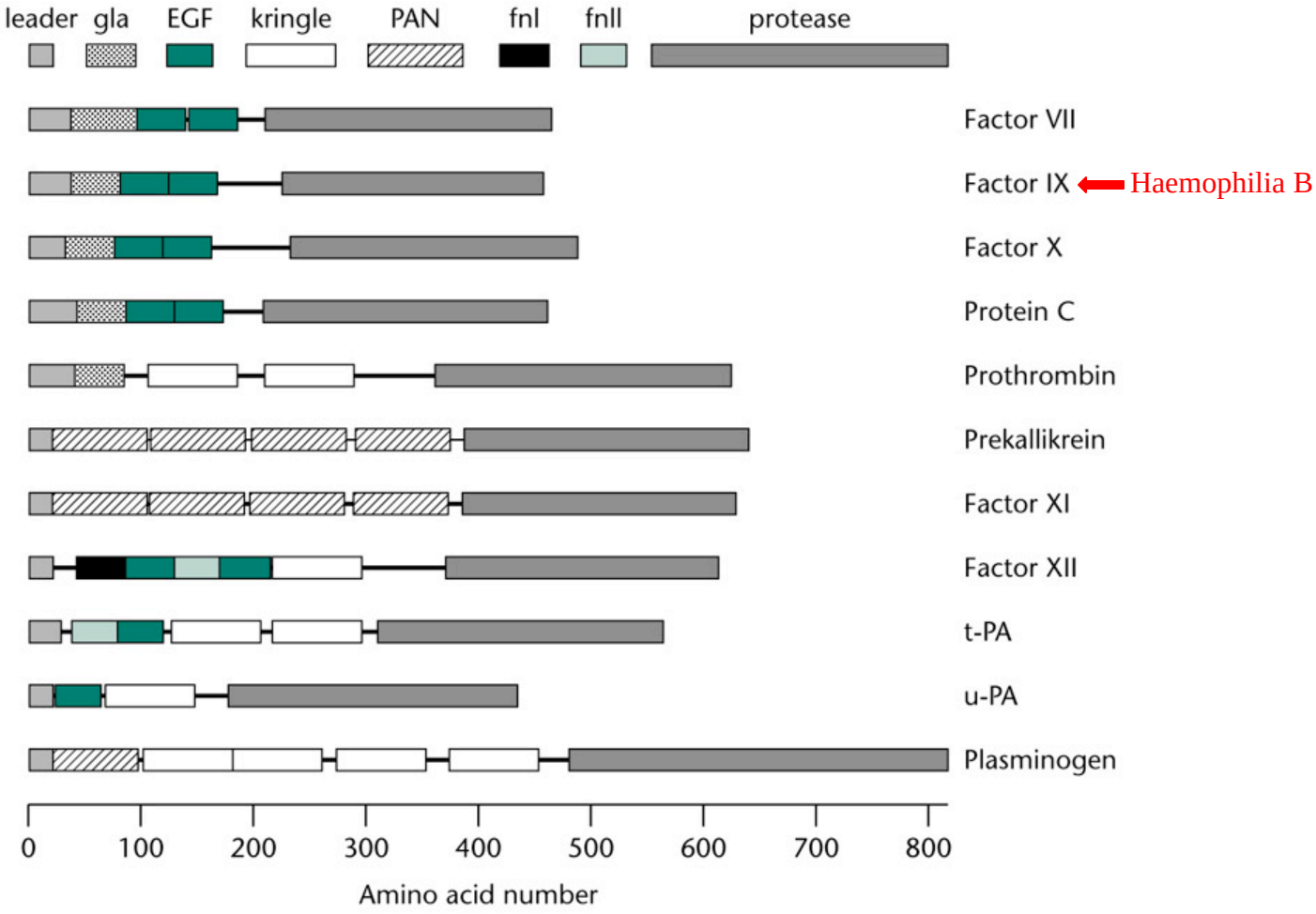
Comparative genomics (contd)

- 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 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) The coffee berry borer (*Hypothenemus hampei*) is a plague that affects coffee crops. *H. hampei* is an exceptional organism, because it consumes so much caffeine that it would kill any other insect - it is exposed to the equivalent of 500 espressos per day. The closely related species, *Hypothenemus eruditus*, aka the bark beetle doesn't like coffee and dies rather quickly when fed ground up coffee beans.

a) (4 points) Starbucks has funded your lab to identify what allows these bugs to eat the coffee that they want to sell you. If you are successful and determine how these bugs tolerate caffeine, it might be possible to develop a specific treatment that blocks their destructive ability without spraying large amounts of toxic pesticides on the coffee crop. You decide to first develop a high quality, draft genome sequence from both *H. hampei* and *H. eruditus* to identify genes that may be different between them and could be responsible for the caffeine tolerance of *H. hampei*. **Outline how you will develop the genome sequences and be sure to mention what will challenges you will face and how you will overcome them**.

In order to produce high quality draft genome sequences from both species of bugs, you will want to use nextgen sequencing of the DNA from each to a high depth of coverage, perhaps with both Illumina and 454 sequencing to facilitate the assemblies. The main challenges you will face in assembling the sequences could be an unexpectedly high amount of repeated sequences which you will address by using 2 different methods and sequencing to high depth.

1b) (4 points) Your comparative sequence analysis has not revealed any significant differences in gene sequences that might be related to caffeine metabolism, but the genomes are fragmented. It might be a good idea to actually finish these sequences to high quality (Starbucks has enough money to pay). **One important factor in finishing is the development of a good quality, long range map to order the various contigs and scaffolds that your analysis in 1a produced.** It is up to you to decide how to produce a good map of the genome to enable a high quality assembly (i.e., better than what you obtained in 1a). **Briefly outline what approach you would take, indicating what sorts of markers you might use for the mapping**.

Since you only got a draft genome from 1a, it would be a good idea to finish the genomic sequences. There are a few good approaches that you might employ. One would be to create BAC libraries from both organisms, sequence a large number of BAC end sequences (sequence tagged connectors) and use these to aid in the assembly. This should allow you to produce a good quality map, assemble the sequence, and close gaps. Another approach would be to generate a large number of EST sequences and then map these to either the BAC libraries, or to the genome using HAPPY mapping. An adequate, although less-effective method would be to create radiation hybrid panels and map the EST or STC markers to these.

2. (10 points) Warren Fahy wrote a book called "Fragment" where he describes the bizarre creatures on a Pacific island that have evolved in isolation for ~600 million years. Although the book was sold as science fiction, Henders island actually exists, as do the dangerous, predatory creatures that live there. The U.S. military has quarantined the island since its discovery instead of nuking it as the book described because they want to create weapons from the organisms. Your magazine has received information and samples (including live cells) from 30 species of Henders island animals sent by a concerned whistleblower. The magazine wants to produce a detailed expose of what is going on at Henders island but also wishes to characterize and publish a thorough analysis of the animals since it is unlikely the military will ever even acknowledge their existence.

2a) (5 points) Since the creatures are so bizarre, the first thing to do is for you to figure out which ones are most related to each other and how all of them are related to other living organisms. **How might you go about building an evolutionary tree for the Henders island creatures and figuring out what organisms on Earth they are most related to ?**

This is along the lines of the Lindblad-Toh paper. You will want to generate draft genome sequences from each of the 30 species, and then compare these with each other to determine which ones are the most closely related. Comparing these sequences with each other will enable you to deduce which regions have undergone purifying selection and; therefore, might be most likely to give you clues about why these animals are so different from other organisms. Comparing these genomes with the GENBANK database will identify the closest relatives, which can also be used in tree building to derive an evolutionary tree.

2b) (5 points) Intriguingly, the animals turn out to be most related to a type of shrimp called Mantis shrimp, that comprise the order Stomatopoda. Mantis shrimps are fearsome predators that typically do not reach sizes larger than 12". Perhaps the 600 million year isolation of the Henders island population has allowed an entire ecosystem to be derived from a single Order of crustaceans. Like Mantis shrimps, Henders island predators fall into two general categories: "smashers" that club their prey to death with near supersonic speed and "spearers" that stab their prey with spiny appendages. It turns out that two apparent species of animals, disk ants and Henders wasps, are closely related by DNA sequence, but look and behave entirely different; disk ants are smashers whereas Henders wasps are spearers. **How would you go about identifying genes that are expressed in the clubbing appendage that are different from those expressed in the spearing appendage with a high probability of not missing any?**

I would isolate RNA from the appendages, perform RNA-seq to high depth and compare the sequences you obtain between the wasps and the disk ant appendages. Performing RNA-seq to high depth will give you many thousands of tags for each mRNA and should allow you to see all types of sequences that are present. A less informative approach that might miss rare sequences would be microarray. If you go this route, you will also need to obtain or generate microarrays.

3. (8 points) Luckily, salt water is toxic for Henders island creatures, which has kept them from spreading throughout the world. This is surprising since Mantis shrimp (the closest relative to Henders island fauna) are only found in salt water.

3a) (4 points) Assume that you can isolate nephridia (kidney-like organs that excrete salt) from both Mantis shrimp and disk ants (the most dangerous of the Henders island predators). Your eager intern, Amanda, hypothesizes that Mantis shrimp can eliminate excess salt, whereas the Henders island animals cannot, thus, the salt toxicity. Moreover, she hypothesizes that a copy number variation in the genome is responsible for the phenotype observed. **How might you determine whether a copy number variation anywhere in the genome is responsible for the differential sensitivity of Mantis shrimp and disk ants to salt water?**

The approach to follow here would parallel the Redon paper. The simplest approach would require microarrays that could be used for both Mantis shrimp and disk ants. Then you could simply compare the hybridization patterns and determine whether there are any regions that appear to be over- or under-represented in one genome vs. the other. Whole genome sequencing would be a less desirable approach since it could be difficult to do this in a sufficiently quantitative manner to identify CNV regions. Once you identify a candidate region, you might want to add it back to the disk ants and see whether they are now tolerant of salt water. However, since disk ants are dangerous, you might not actually want to do this experiment in whole organisms….

3b) (4 points) Another very interesting feature of Henders island fauna is their incredible speed. They strike quickly, can move from place to place very fast and are largely resistant to the effects of usual weapons. One hypothesis is that the Henders island animals have a super-efficient metabolism that produces much more energy than typical animals. Amanda (who is full of ideas) hypothesizes that the gut microbes are responsible for the incredible metabolic efficiency of animals such as disk ants. **Describe how would you determine how many kinds of bacteria are found in disk ant digestive organs? How would you relate these to known species of bacteria?**

The most desirable approach would be something like the Venter sequencing paper. You will want to isolate material from the digestive organs and perform total shotgun sequencing using nextgen sequencing, then assemble these sequences into genomes. Once you have the best assembly of the microbial genomes, you can compare these with each other and with GENBANK to determine how closely related the microbes are to each other and to known organisms.

4. (4 points) Disk ants are disks ~1 foot in diameter that can rapidly scurry around on either side and roll even faster when on edge. They have a very unpleasant property that interests you quite a bit - when you bash them to bits, they lay around for a while, but then regenerate within a few hours unless you burn the pieces. You recall that a Nature paper once showed that mouse white blood cells can sometimes become pluripotent stem cells when stressed and give rise to entire organisms when injected into mice. Perhaps this is a clue about what is happening with the disk ants. Amanda suggests that take some of the disk ant cells, subject these to stress, then observe whether they can, in fact, become stem-cell like. **How would you perform whole transcriptome analysis to identify which RNAs were altered in expression by exposing the cells to a stress such as salt water, acid or forcing them through a very narrow tube? What is a critical thing that you will need to do this analysis?** Assume that you are able to identify whether cells are stem cells or not by observation.

This is yet another application of whole transcriptome analysis so you will want to use RNA-seq again. You will need to have completed the disk ant genome before the RNA-seq will be useful since you need to map reads to the genome in order to be sure that the sequences you are obtaining actually come from disk ant material. If you chose to use microarray analysis to do transcriptomics, you will absolutely need microarrays.

5. (5 points) One of the crew, Wabadummy, on the boat that brought you to Henders island was both brave and foolish - he decided to make sashimi out of disk ants before you learned about their regenerative properties. A few days after eating the sashimi, Wabadummy became very energetic and appeared to have "super strength". He lifted a lifeboat off its support and threw it into the ocean then jumped in and swam around the ship many times at an astonishing rate of speed before he got tired, was apprehended and returned for observation (in the brig). Amanda hypothesizes that gut bacteria from the disk ant gave Wabadummy "super strength" like they do to the disk ants.

a) (3 points) **How would you develop a test that determines whether the major species of gut bacterium *Sehrstark hendersii*, found in disk ants, was present in Wabadummy's intestines?**

You will want to develop a quantitative PCR assay that can rapidly and easily identify whether material from *Sehrstark hendersii* is present in any type of sample. Such an assay requires primers to amplify the material and either Sybr Green to detect the amplicons, or a Taqman type probe to directly identify amplified products. Since you want to identify material from gut bacteria, take a fecal sample from Wabadummy, prepare DNA and use your assay to determine whether S. hendersii DNA is present. Sequencing the entire gut genome would be overkill.

b) (2 points) Because you are a thorough scientist, you wonder whether any other members of the crew have acquired Henders island bacteria in their intestines. **How could you test this in a cost effective manner and justify why you chose the method you did?**

You are going to first design PCR assays that detect the bacterial groups that you identified in question 3b. This is a good idea because you want to identify many bacterial species in a cost-effective manner, rather than sequence the entire gut bacteria genomes of the entire crew.

