

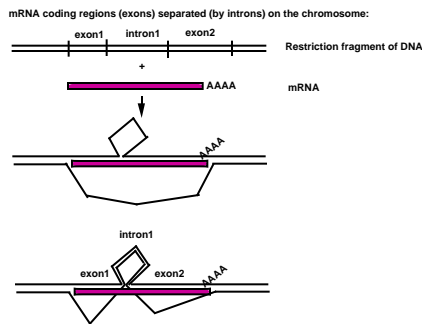
Fine Structure and Analysis of Eukaryotic Genes

Split genes
Multigene families
Functional analysis of eukaryotic genes

Split genes and introns

- The mRNA-coding portion of a gene can be split by DNA sequences that do not encode mature mRNA
- Exons** code for mRNA, **introns** are segments of genes that do not encode mRNA.
- Introns are found in most genes in eukaryotes
- Also found in some bacteriophage genes and in some genes in archae

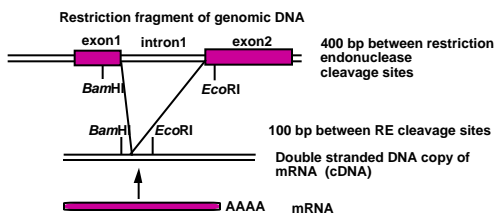
R-loops can reveal introns



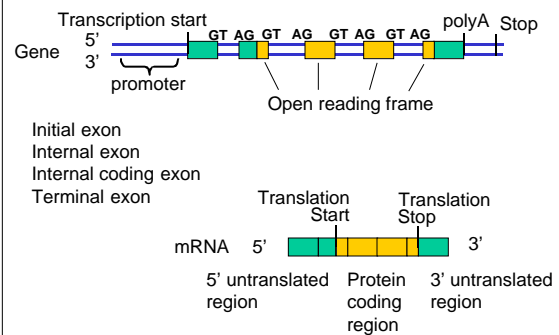
Examples of R-loops in mammalian hemoglobin genes



Comparison of cDNA and genomic clone maps can reveal introns, 1.6.3



Types of exons



Finding exons with computers

- *Ab initio* computation
 - E.g. Genscan: <http://genes.mit.edu/GENSCAN.html>
 - Uses an explicit, sophisticated model of gene structure, splice site properties, etc to predict exons
- Compare cDNA sequence with genomic sequence
 - BLAST2 alignments between cDNA and genomic sequences
 - <http://www.ncbi.nlm.nih.gov/blast/>
 - Better: Use sim4
 - Takes into account terminal redundancy at ends of introns
 - <http://bio.cse.psu.edu>
 - Follow link to "sim4 server in France"

Find exons for *HBB*

- Sequence for human beta-globin gene (*HBB*):
 - Accession number L48217
 - Thalassemia variant
- Sequence for *HBB* mRNA
 - NM_000518
- Retrieve those from GenBank at NCBI (or the course website)
 - <http://www.ncbi.nlm.nih.gov>
 - Get the files in FASTA format
- Run Genscan and BLAST2 sequences

Genscan analysis of *HBB* gene

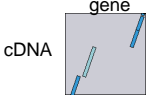
GENSCAN 1.0 Date run: 8-Sep-100 Time: 11:29:36
 Sequence gi : 1827 bp : 41.54% C+G : Isochore 1 (0 - 43 C+G%)
 Parameter matrix: HumanIso.smat
 Predicted genes/exons:

Gn.Ex	Type	S	.Begin	...End	.Len	Fr	Ph	I/Ac	Do/T	CodRg	P....	Tscr...
1.01	Init	+	217	308	92	0	2	103	77	136	0.987	14.01
1.02	Intr	+	439	661	223	1	1	100	96	217	0.999	20.91
1.03	Term	+	1512	1640	129	2	0	116	43	119	0.862	7.40
1.04	PlyA	+	1667	1672	6							-1.95

Predicted peptide sequence(s):
 >gi|GENSCAN_predicted_peptide_1|147_aa
 MVHLTPEEKSAVTALWGKLVNVEVGGELGRLLVYIPWTQRFESFGDLSTPDVAMGNPK
 VKAHGKVLGAFSDGLAHLNLKGTFTALSELHCDKLVDPENFRLLGNLVLCVLAHFHG
 KEFTPPVQAAYQKVVAVANALAHKYH

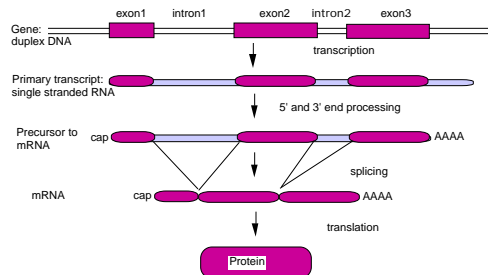
BLAST2: *HBB* gene vs. cDNA

gene
 cDNA

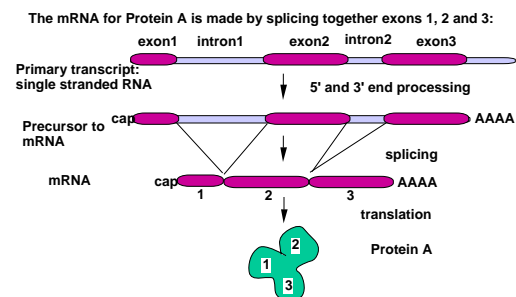


Score = 275 bits (143), Expect = 1e-71
 Identities = 143/143 (100%), Positives = 143/143 (100%)
 Query: 16atttctctctgtgacacaactgtgttctactagcaacctcaaacagacacccatggtgacc 226
 Sbjct: 1acatttctctgtgacacaactgtgttctactagcaacctcaaacagacacccatggtgacc 60
 hemoglobin, beta 1 H V H
 Query: 22tactctgagggagaagctgccttactgccctgtggggcaagtgaaactggtgaaag 286
 Sbjct: 61 tgactctgagggagaagctgccttactgccctgtggggcaagtgaaactggtgaaag 120
 hemoglobin, beta 4 L T P E E K S A V T A L W G K V N V D E
 Query: 287 ttggtgtgagggccctggggcagg 309
 Sbjct: 121 ttggtgtgagggccctggggcagg 143
 hemoglobin, beta 24 V G G E A L G R

Introns are removed by splicing RNA precursors

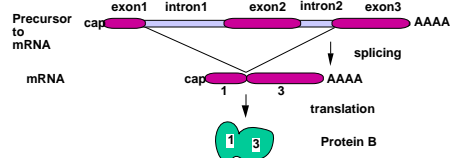


Alternative splicing can generate multiple polypeptides from a single gene

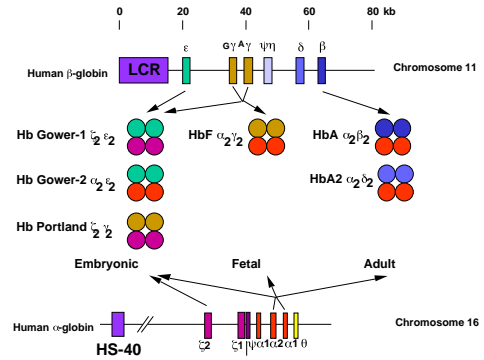


Alternative splicing can generate multiple polypeptides from a single gene, part 2

Or, by an alternative pathway of splicing that skips over exon2, Protein B can be made:

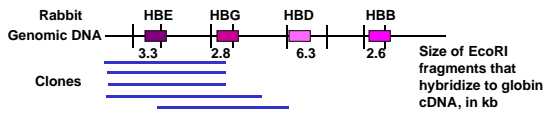
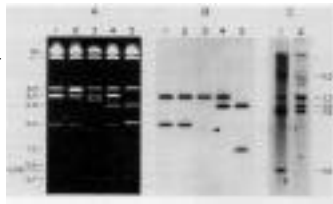


Multigene families, e.g. encoding hemoglobin

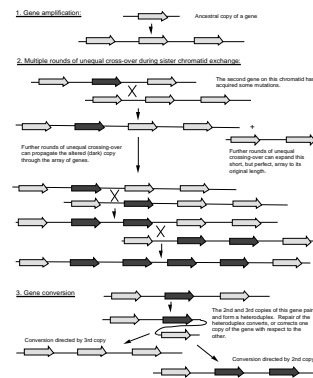


Blot-hybridization analysis showing multiple beta-like globin genes in mammals

A: clones, gel
B: clones, blot-hybridization
C: genomic DNA, blot-hybridization



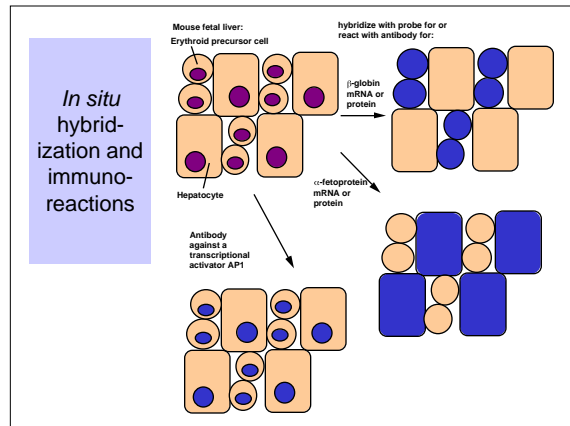
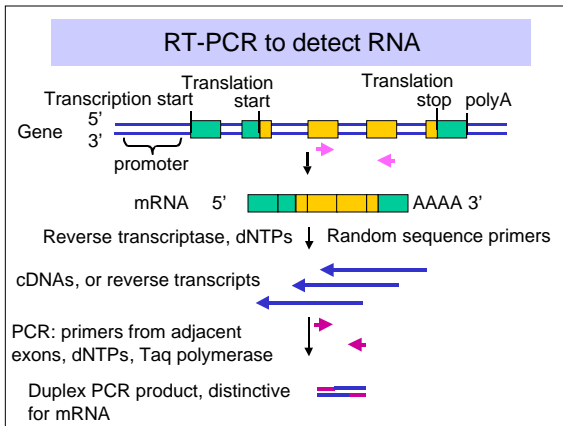
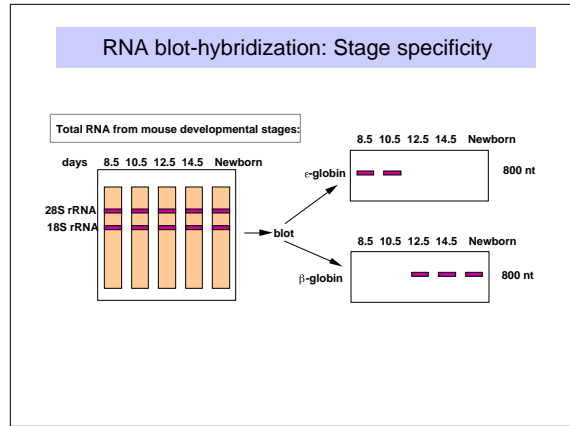
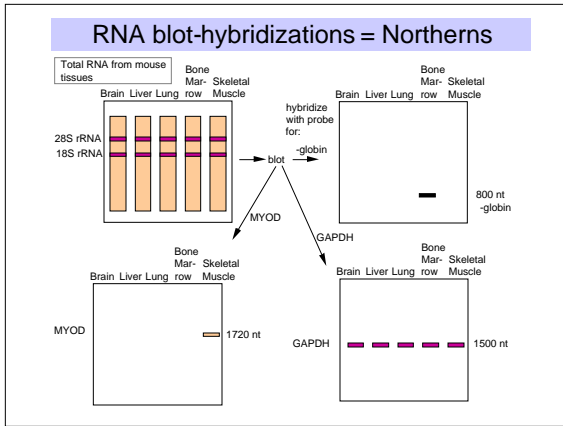
Maintaining sequence similarity in gene families: Unequal cross-overs and gene conversions



Functional analysis of isolated genes

Gene Expression: where and how much?

- A gene is *expressed* when a functional product is made from it.
- One wants to know many things about how a gene is expressed, e.g.
 - In which tissues?
 - At what developmental stages?
 - In response to which environmental conditions?
 - At which stages of the cell cycle?
 - How much product is made?

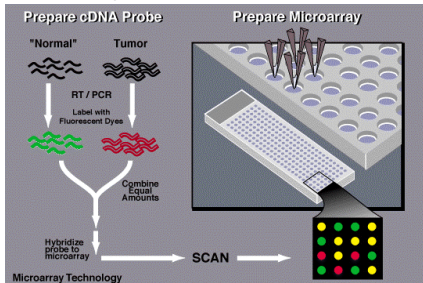


- ### Sequence everything, find function later
- Determine the sequence of hundreds of thousands of cDNA clones from libraries constructed from many different tissues and stages of development of organism of interest.
 - Initially, the sequences are partials, and are referred to as expressed sequence tags (ESTs).
 - Use these cDNAs in high-throughput screening and testing, e.g. expression microarrays (next presentation).

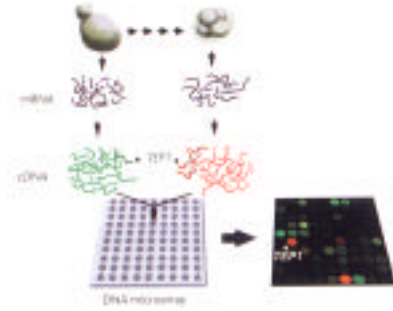
- ### Massively parallel screening of high-density chip arrays
- Once the sequence of an entire genome has been determined, a diagnostic sequence can be generated for **all** the genes.
 - Synthesize this diagnostic sequence (a tag) for each gene on a high-density array on a chip, e.g. 6000 to 20,000 gene tags per chip.
 - Hybridize the chip with labeled cDNA from each of the cellular states being examined.
 - Measure the level of hybridization signal from each gene under each state.
 - Identify the genes whose expression level differs in each state. The genes are already available.

Hybridization of RNA to "Gene chips"

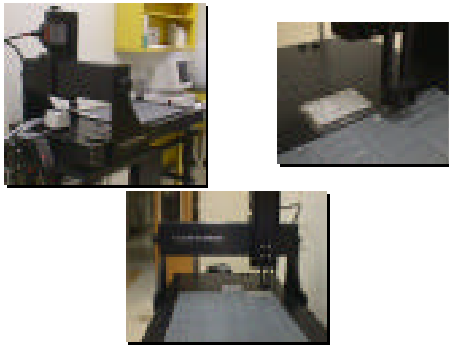
Gene chip = high density microarray of sequences from many (all) genes of an organism



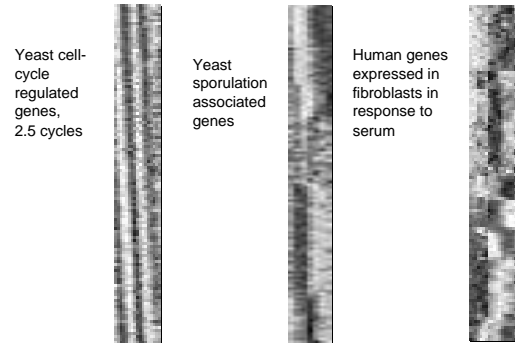
Expression profiling using microarrays



PSU's microarray spotting robot



Find clusters of co-regulated genes



Spellman et al. (1998) Mol. Biol. Cell 9:3273; Chu et al. (1998) Science 282:699; Iyer et al. (1999) Science 283:83.

Search the databases

- What can be learned from the DNA sequence of a novel gene or polypeptide?
- Many metabolic functions are carried out by proteins conserved from bacteria or yeast to humans - one may find a homolog with a known function.
- Many sequence motifs are associated with a specific biochemical function (e.g. kinase, ATPase). A match to such a motif identifies a potential class of reactions for the novel polypeptide.

Databases, cont'd

- One may find a match to other genes with no known function, but their pattern of expression may be known.
- Types of databases:
 - Whole and partial genomic DNA sequences
 - Partial cDNAs from tissues (ESTs = expressed sequence tags)
 - Databases on gene expression
 - Genetic maps

Express the protein product

- Express the protein in large amounts
 - In bacteria
 - In mammalian cells
 - In insect cells (baculovirus vectors)
- Purify it
- Assay for various enzymatic or other activities, guided by (e.g.)
 - The way you screened for the clone
 - Sequence matches

Phenotype of directed mutation

- Mutate the gene in the organism of interest, and then test for a phenotype
- Gain of function
 - Over-expression
 - Ectopic expression (where normally is silent)
- Loss of function
 - Knock-out expression of the endogenous gene (homologous recombination, antisense)
 - Express dominant negative alleles
 - Conditional loss-of-function, e.g. knock-out by recombination only in selected tissues

Localization on a gene map

- E.g., use gene-specific probes for *in situ* hybridizations to mitotic chromosomes. Align the hybridization pattern with the banding pattern
- Are there any previously mapped genes in this region that provide some insight into your gene?