

Chapter 10:

Bioinformatic approaches to ribonucleic acid (RNA)

Learning objectives

- describe the major categories of coding and noncoding RNA;
- compare and contrast techniques for measuring steady-state RNA levels; and
- compare and contrast the use of microarrays and RNA-seq for measuring mRNA levels.

Outline

Introduction to RNA

Noncoding RNA

Rfam; tRNA; ribosomal RNA; small nuclear RNA; small nucleolar RNA; microRNA; short interfering RNA; long noncoding RNA; UCSC

Introduction to messenger RNA

mRNA; low- and high-throughput technologies; cDNA libraries; full-length cDNA; BodyMap2, GTEx

Microarrays and RNA-seq

Stage 1: experimental design

Stage 2: RNA preparation and probe preparation

Stage 3: data acquisition

Stage 4: data analysis

Stage 5: biological confirmation

Microarray and RNA-seq Databases

Interpretation of RNA analyses

Relationship between DNA, mRNA, and protein

Pervasive nature of transcription

eQtls: genetic basis of variation in gene expression

Perspective

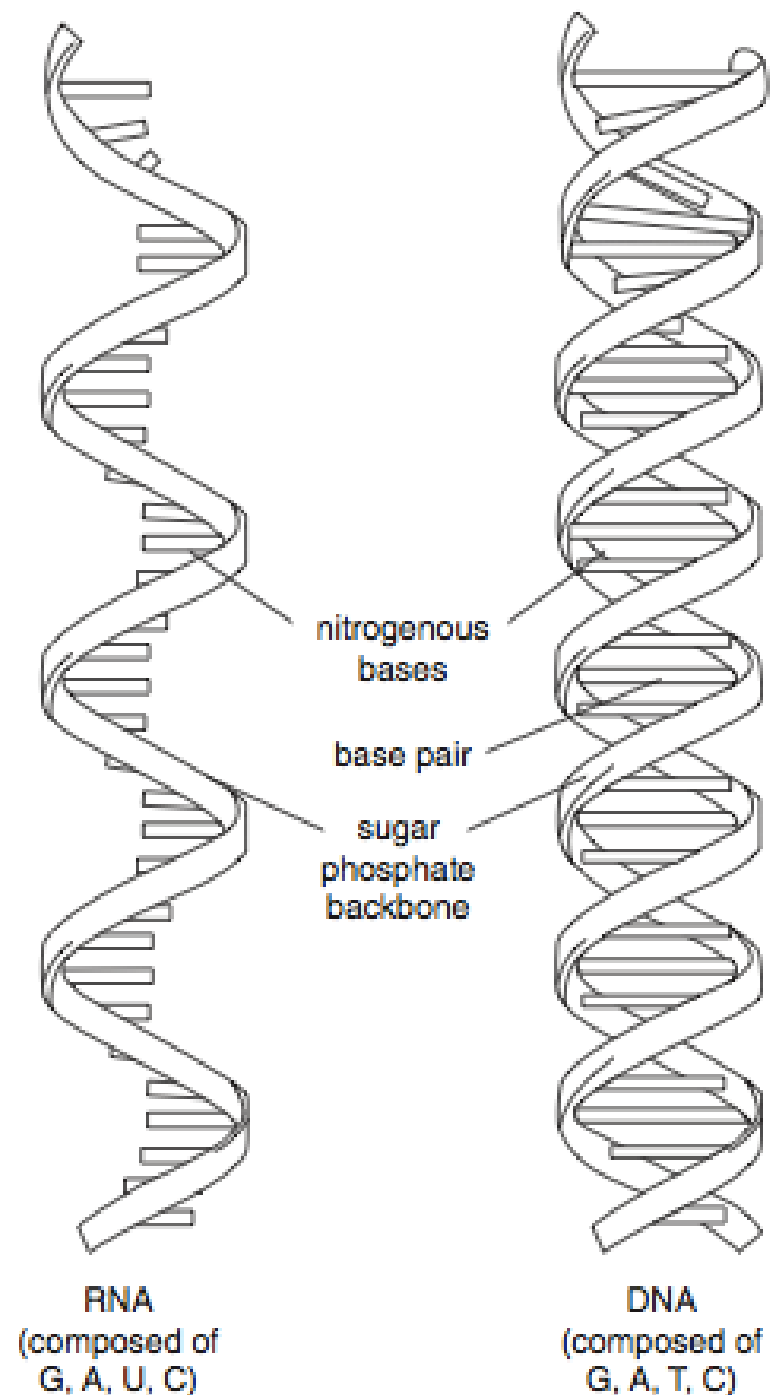
DNA and RNA

In 1953 Crick and Watson discovered the double helical nature of DNA.

DNA: A, G, C, T
adenine, guanine, cytosine, thymidine

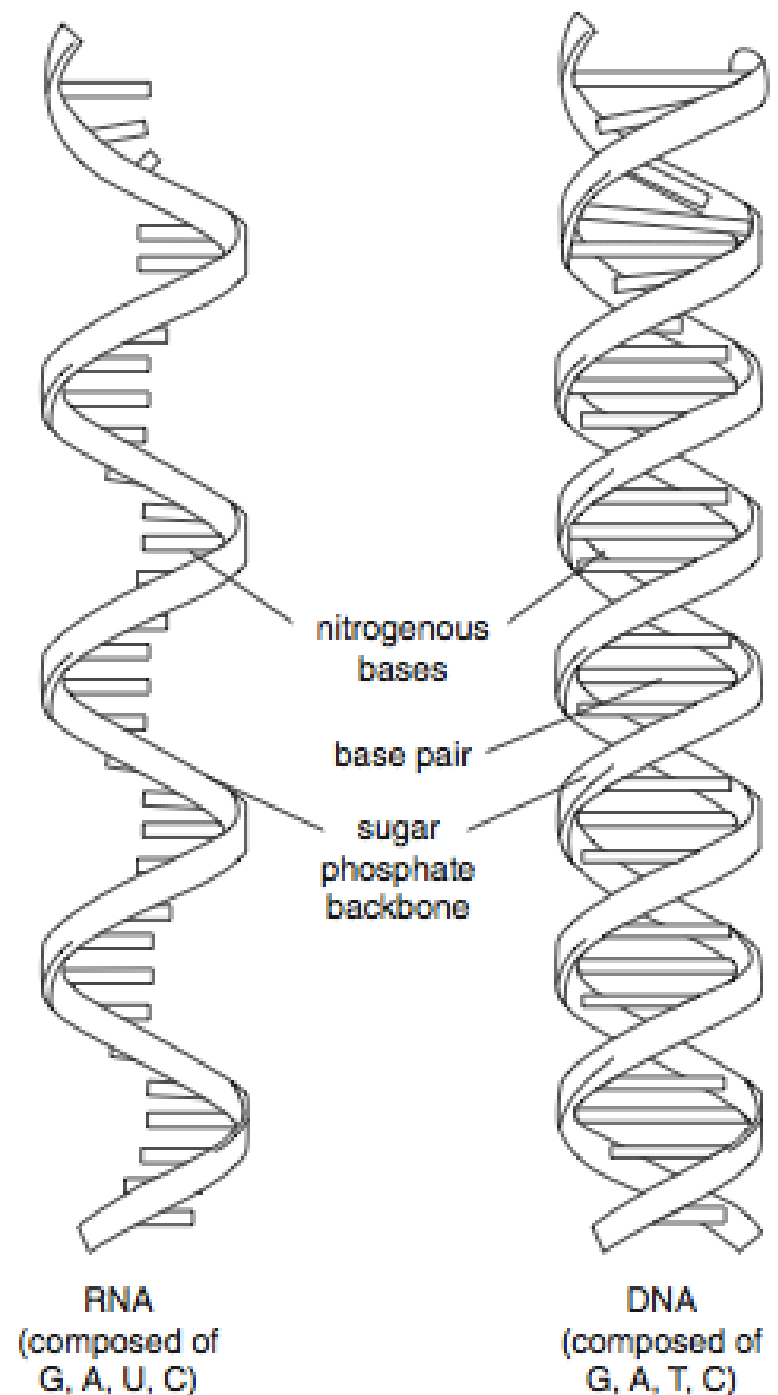
RNA: G, A, U, C
U = uracil

Crick further helped elucidate the nature of the genetic code, proposing the existence of tRNA.

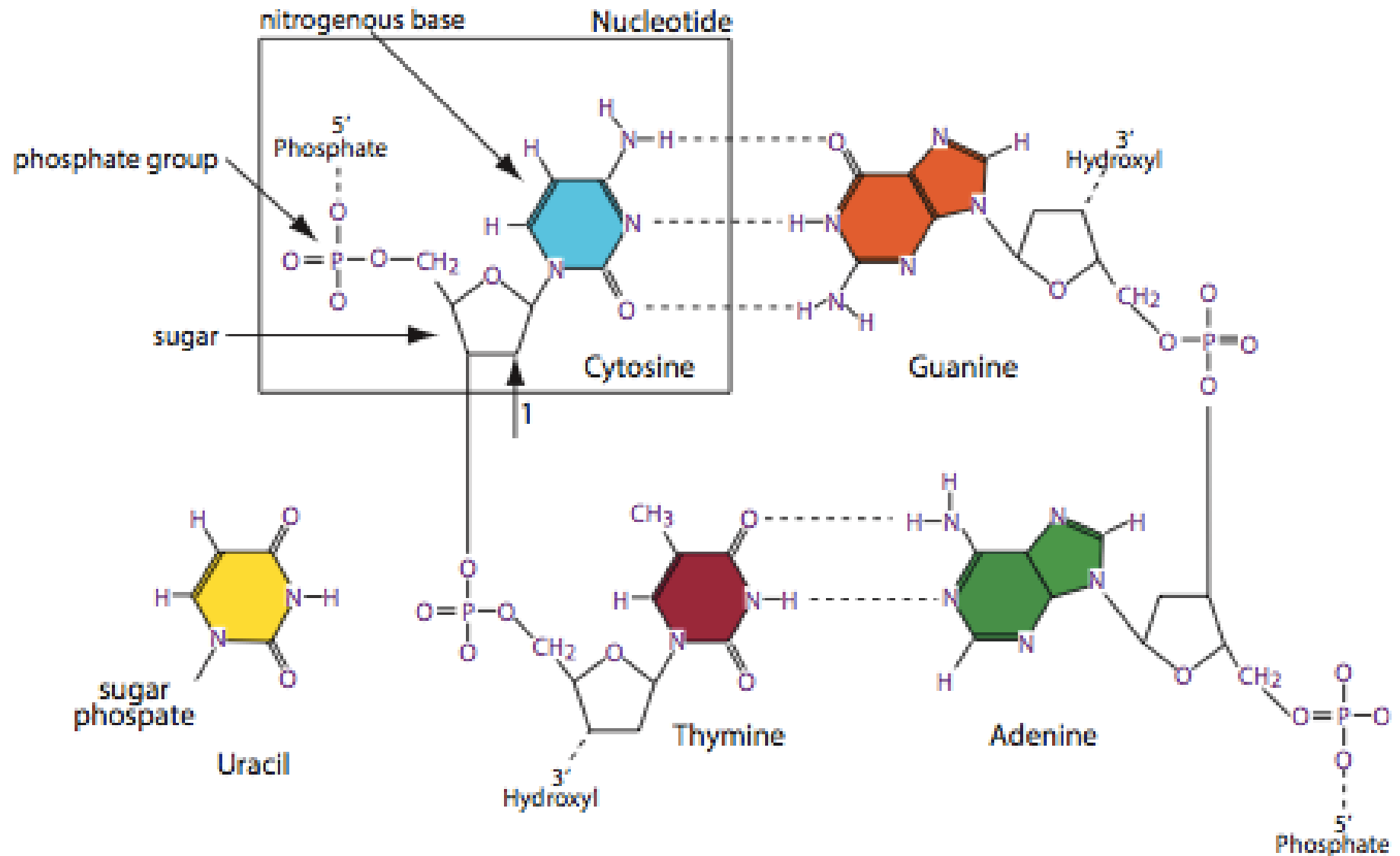


DNA and RNA

Crick (1958) wrote that the central dogma “states that once ‘information’ has passed into protein *it cannot get out again*. In more detail, the transfer of information from nucleic acid to nucleic acid, or from nucleic acid to protein may be possible, but transfer from protein to protein, or from protein to nucleic acid is impossible. Information means here the *precise* determination of sequence, either of bases in the nucleic acid or of amino acid residues in the protein.”



Nucleotide bases



Outline

Introduction to RNA

Noncoding RNA

- Rfam; tRNA; ribosomal RNA; small nuclear RNA; small nucleolar RNA; microRNA; short interfering RNA; long noncoding RNA; UCSC

Introduction to messenger RNA

mRNA; low- and high-throughput technologies; cDNA libraries; full-length cDNA; BodyMap2, GTEx

Microarrays and RNA-seq

Stage 1: experimental design

Stage 2: RNA preparation and probe preparation

Stage 3: data acquisition

Stage 4: data analysis

Stage 5: biological confirmation

Microarray and RNA-seq Databases

Interpretation of RNA analyses

Relationship between DNA, mRNA, and protein

Pervasive nature of transcription

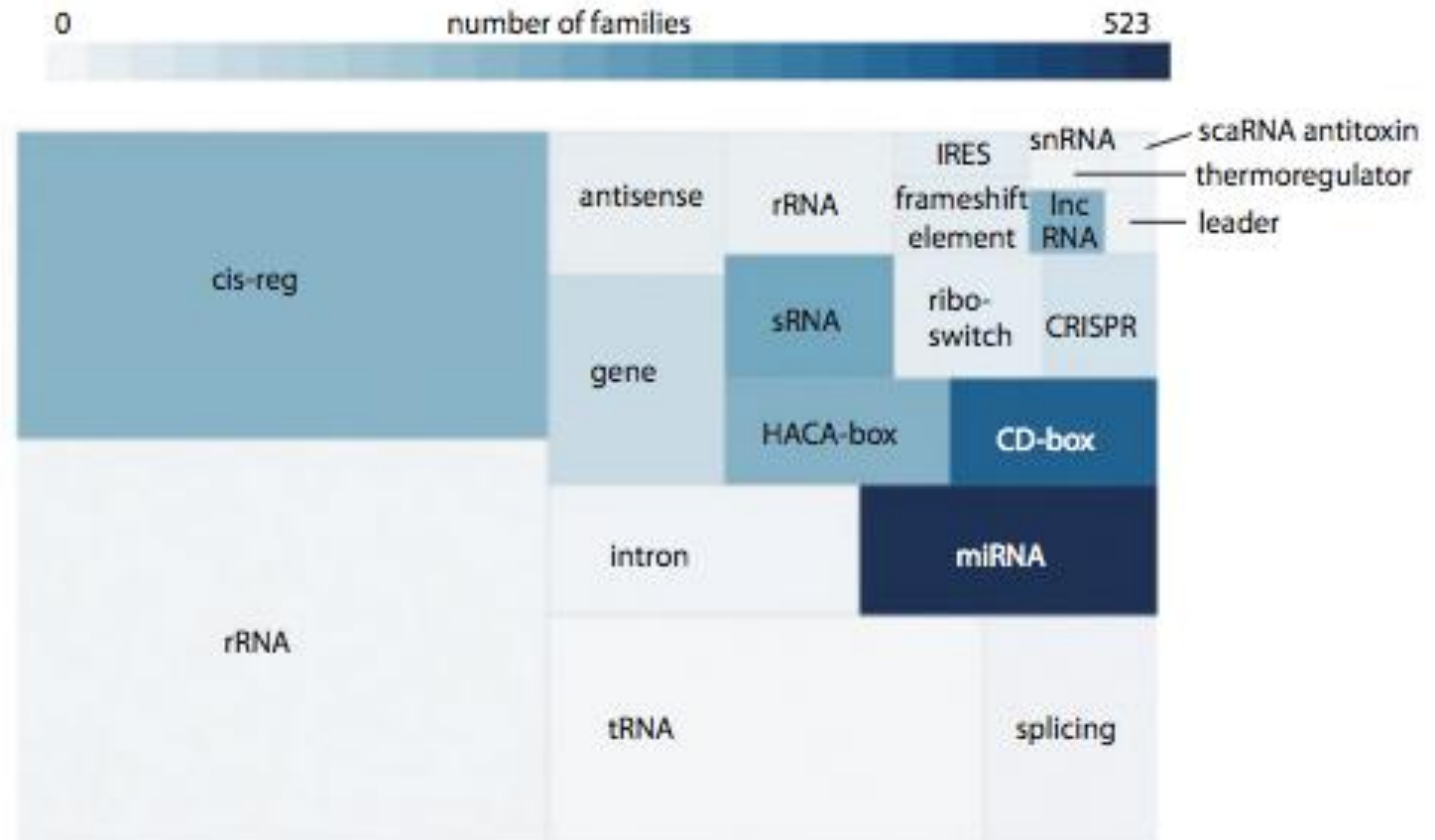
eQtls: genetic basis of variation in gene expression

Perspective

Noncoding RNA

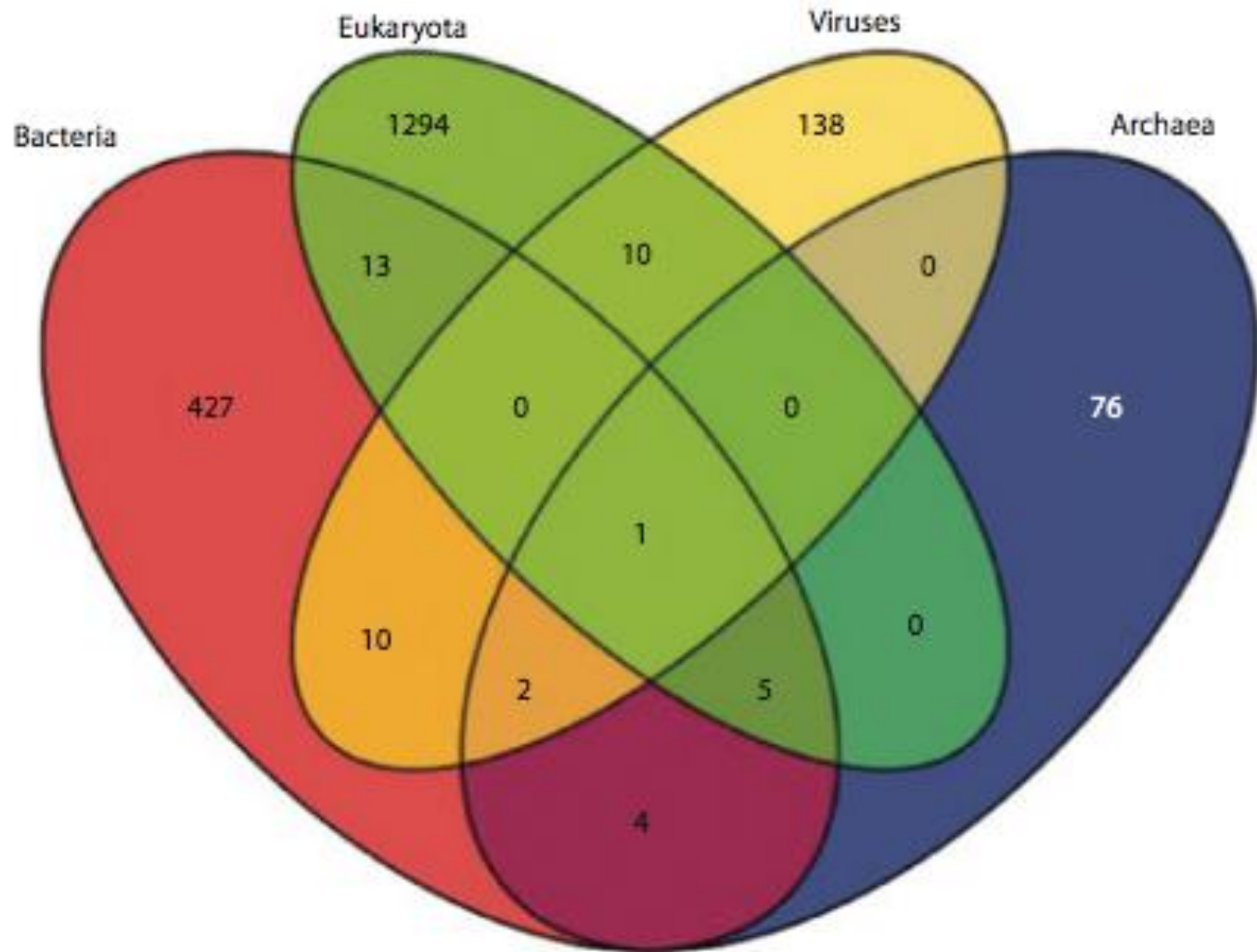
- The main kinds of noncoding RNA are ribosomal RNA (rRNA) and transfer RNA (tRNA)
- *XIST* is an example of a well characterized noncoding RNA gene. It is located in the X inactivation center of the X chromosome and functions in X chromosome inactivation. While males have one copy of the X chromosome (with XY sex chromosomes), females have two copies of which one is inactivated in every diploid cell of mammalian and some other species. *Xist* is expressed from the inactive X and binds to its chromatin, facilitating chromosome inactivation.
- Rfam is a database of noncoding RNA families across the tree of life.

Rfam database



Rfam sequence space and numbers of families

Rfam database



Rfam taxonomic groupings

13 Rfam entries with the largest number of members

Name	Accession	No. full	Ave. len. (full)	Id	Type	Description
5_8S_rRNA	RF00002	376,000	152	69	Gene; rRNA	5.8S ribosomal RNA
tRNA	RF00005	298,000	73	46	Gene; tRNA	tRNA
5S_rRNA	RF00001	229,000	116	60	Gene; rRNA	5S ribosomal RNA
UnaL2	RF00436	101,000	54	78	Cis-reg	UnaL2 LINE 3' element
HIV_POL-1_SL	RF01418	83,000	113	77	Cis-reg	HIV pol-1 stem loop
U6	RF00026	72,000	105	77	Gene; snRNA; splicing	U6 spliceosomal RNA
mtDNA ssA	RF01853	62,000	104	67	Gene; antisense	Mitochondrial DNA control region secondary structure A
Intron_gpl	RF00028	60,000	365	36	Intron	Group I catalytic intron
Intron_gpll	RF00029	51,000	87	54	Intron	Group II catalytic intron
Hammerhead_1	RF00163	49,000	59	70	Gene; ribozyme	Hammerhead ribozyme (type I)
RRE	RF00036	44,000	337	97	Cis-reg	HIV Rev response element
HIV_GSL3	RF00376	39,000	84	82	Cis-reg	HIV gag stem loop 3 (GSL3)
SNORA7	RF00409	26,000	140	79	Gene; snRNA; snoRNA; HACA-box	Small nucleolar RNA SNORA7

No. full: number of members of the Rfam family

Id: average percent identity of the full alignments

Transfer RNA

Transfer RNA molecules carry a specific amino acid and match it to its corresponding codon on an mRNA during protein synthesis. tRNAs occur in 20 amino acid acceptor groups corresponding to the 20 amino acids specified in the genetic code.

tRNA forms a **structure** consisting of about 70–90 nucleotides folded into a characteristic **cloverleaf**. Key features of this structure include a **D loop**, an **anti-codon loop** which is responsible for **recognizing messenger RNA codons**, a **T loop**, and a 3' end to which **aminoacyl tRNA synthetases** attach the appropriate amino acid **specific for each tRNA**.

Noncoding RNA families in the Rfam database assigned to human chromosome 21

Family	Start	End	Bits score
<u>tRNA</u>	9,734,391	9,734,325	31.22
<u>RSV RNA</u>	9,990,192	9,989,909	36.74
<u>RSV RNA</u>	10,142,311	10,142,595	36.83
<u>Metazoa SRP</u>	10,380,661	10,380,378	122.56
<u>SNORA70</u>	10,385,953	10,386,047	42.25
<u>tRNA</u>	10,492,972	10,492,907	26.05
<u>tRNA</u>	10,493,037	10,492,973	37.46
<u>mir-548</u>	11,052,015	11,051,932	82.85
<u>U6</u>	14,419,904	14,420,010	66.41
<u>U6</u>	14,993,898	14,994,004	76.41
<u>U6</u>	15,340,916	15,340,810	63.69
<u>5S rRNA</u>	15,443,192	15,443,307	42.69
<u>pRNA</u>	15,448,359	15,448,271	68.52
<u>U6</u>	16,986,602	16,986,708	75.19
<u>U6</u>	17,407,829	17,407,733	41.70

<u>SNORD74</u>	17,657,089	17,657,017	59.88
<u>mir-10</u>	17,911,414	17,911,485	69.08
<u>let-7</u>	17,912,152	17,912,227	62.65
<u>lin-4</u>	17,962,567	17,962,636	76.22
<u>U1</u>	18,091,317	18,091,476	91.09
<u>U6</u>	18,803,865	18,803,965	62.92
<u>tRNA</u>	18,827,177	18,827,107	63.87
<u>Metazoa SRP</u>	18,878,771	18,879,046	64.51
<u>Y RNA</u>	18,899,565	18,899,458	41.00
<u>Y RNA</u>	18,949,116	18,949,224	40.22
<u>RSV RNA</u>	19,938,102	19,937,818	72.79
<u>U1</u>	20,717,465	20,717,629	93.53
<u>U6</u>	21,728,164	21,728,060	49.35
<u>7SK</u>	21,728,965	21,729,208	75.15
<u>mir-492</u>	21,798,181	21,798,066	40.04
<u>U4</u>	23,577,511	23,577,651	73.90
<u>U2</u>	24,654,231	24,654,058	62.66

Identification of tRNAs using the tRNAscan-SE server

Isotype / Anticodon Counts:

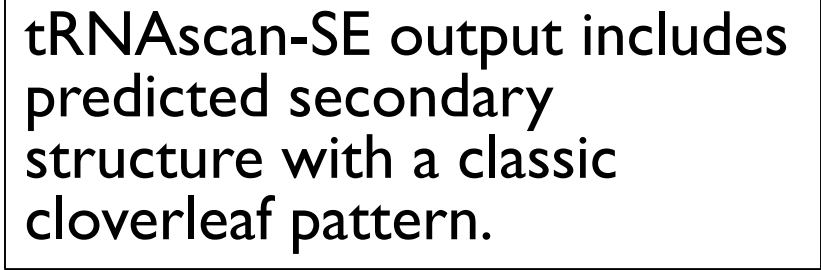
Ala	: 0	AGC:	GGC:	CGC:	TGC:		
Gly	: 1	ACC:	GCC: 1	CCC:	TCC:		
Pro	: 0	AGG:	GGG:	CGG:	TGG:		
Thr	: 0	AGT:	GGT:	CGT:	TGT:		
Val	: 0	AAC:	GAC:	CAC:	TAC:		
Ser	: 0	AGA:	GGA:	CGA:	TGA:	ACT:	GCT:
Arg	: 0	ACG:	GCG:	CCG:	TCG:	CCT:	TCT:
Leu	: 0	AAG:	GAG:	CAG:	TAG:	CAA:	TAA:
Phe	: 0	AAA:	GAA:				
Asn	: 0	ATT:	GTT:				
Lys	: 0			CTT:	TTT:		
Asp	: 0	ATC:	GTC:				
Glu	: 0			CTC:	TTC:		
His	: 0	ATG:	GTG:				
Gln	: 0			CTG:	TTG:		
Ile	: 0	AAT:	GAT:		TAT:		
Met	: 0			CAT:			
Tyr	: 0	ATA:	GTA:				
Supres:	0			CTA:	TTA:		
Cys	: 0	ACA:	GCA:				
Trp	: 0			CCA:			
SecCys:	0				TCA:		

Your-seq.trnal (1-71) Length: 71 bp

Type: Gly Anticodon: GCC at 33-35 (33-35) Score: 71.03

Seq: GCATGGGTGGTTCAGTGGTAGAATTCTCGCCTGCCACGCGGGAGGCCCGGGTTCGATTCCCGGCCCATGCA
Str: >>>>>>...>>>>.....<<<<.>>>>.....<<<<....>>>>.....<<<<<<<<<<

Input 71 base pairs of DNA. tRNAscan-SE output includes anticodon counts and predicted secondary structure.



RNA structure prediction based on the minimum free energy of folding

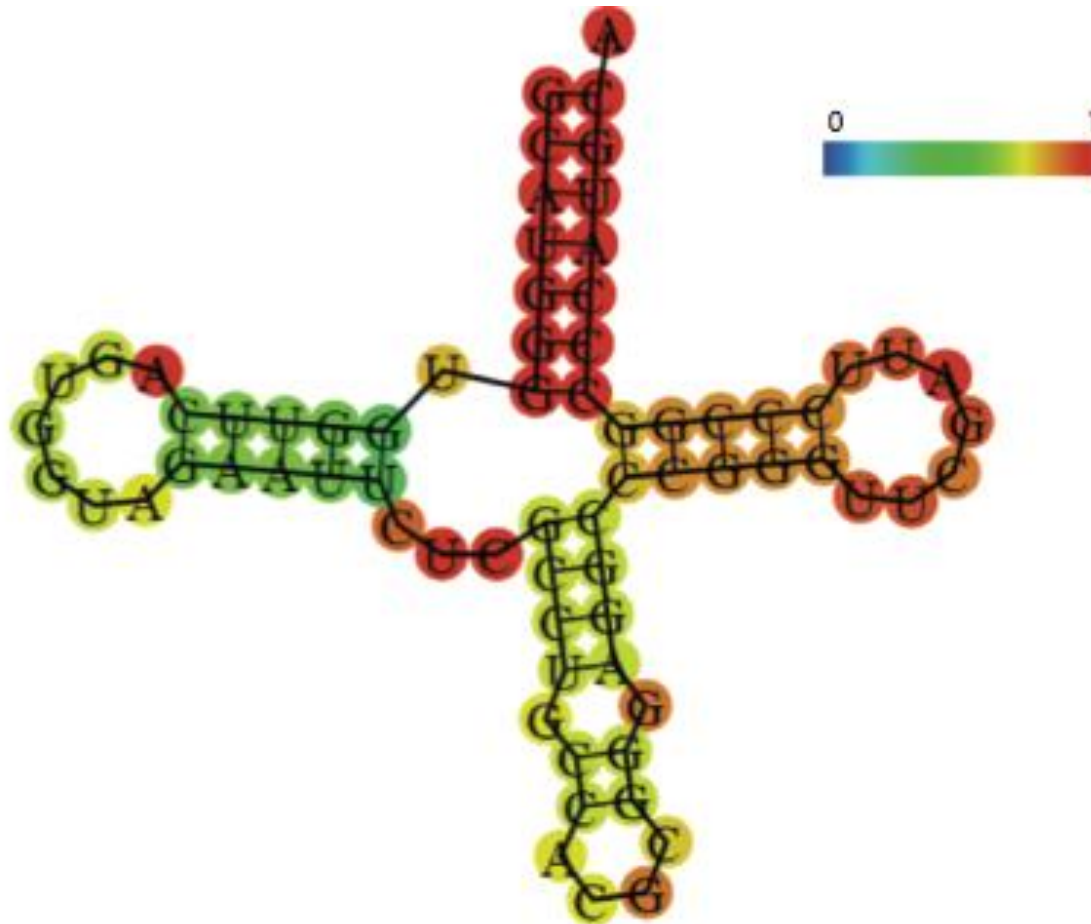
Minimum free energy prediction



A thermodynamic approach to tRNA prediction is implemented in programs such as the Vienna RNA package.
<https://www.tbi.univie.ac.at/RNA/>

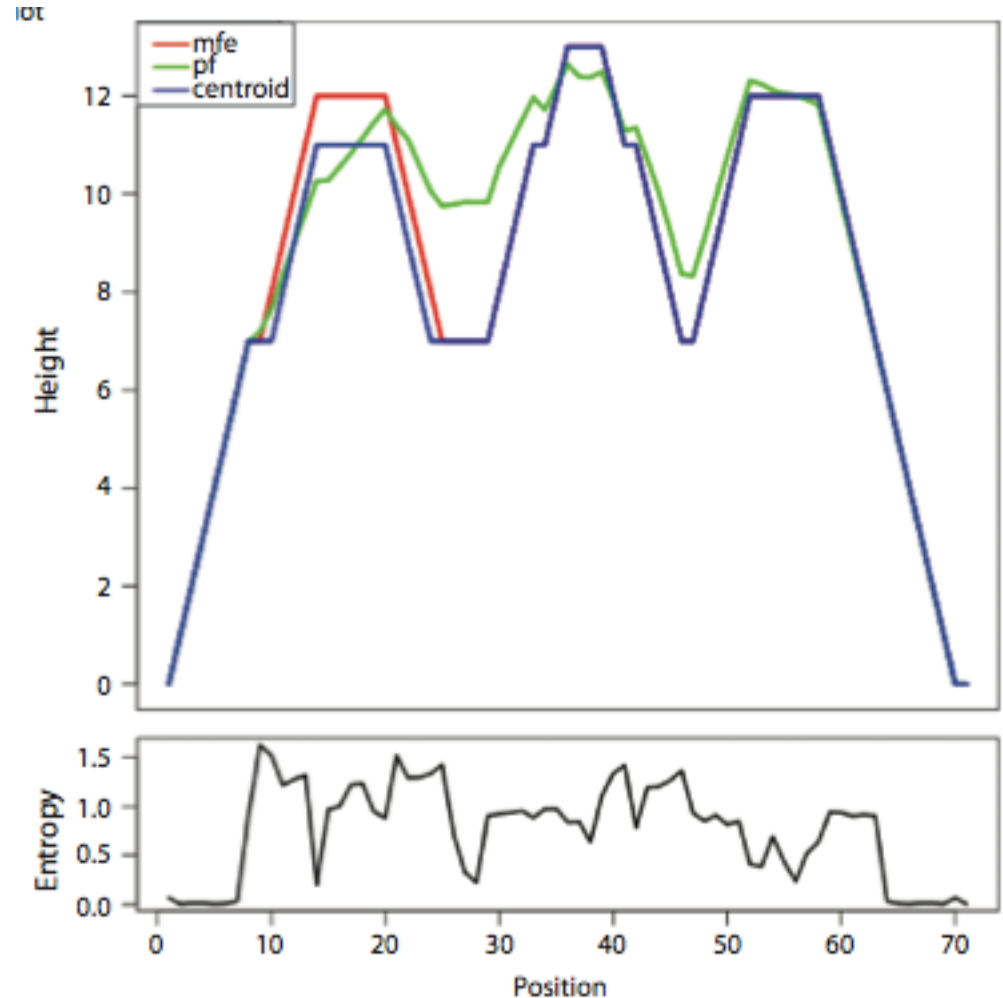
RNA structure prediction based on the minimum free energy of folding

Minimum free energy secondary structure



tRNA structure prediction based on the minimum free energy of folding

Mountain plot



Minimum free energy is plotted based on entropy measurements (y-axis) versus sequence position (x-axis).

Number of tRNA genes in selected organisms

TABLE 10.2 Summary of the number of tRNA genes in selected organisms. The “other” category refers to selenocysteine tRNAs (TCA), suppressor tRNAs (CTA,TTA), or tRNAs with undetermined or unknown isotypes. Additionally, some organisms have tRNAs with introns (e.g., human, 32; *P. falciparum*, 1; *Arabidopsis*, 83).

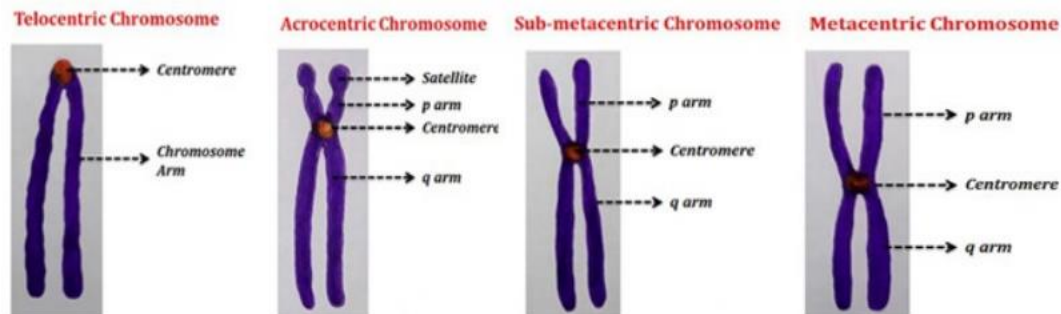
Organism	Common name	No. tRNAs decoding the 20 amino acids	No. predicted pseudogenes	Other	Total
<i>Homo sapiens</i>	Human	506	110	9	625
<i>Pan troglodytes</i>	Chimpanzee	456	0	3	459
<i>Mus musculus</i>	Mouse	432	0	3	435
<i>Canis familiaris</i>	Dog (Canfam1)	898	0	8	906
<i>Drosophila melanogaster</i>	Fruit fly	298	4	2	304
<i>Saccharomyces cerevisiae</i>	Baker's yeast	286	6	3	295
<i>Arabidopsis thaliana</i>	Plant	630	8	1	639
<i>Plasmodium falciparum</i>	Malaria parasite	35	0	0	35
<i>Methanococcus jannaschii</i>	Archaeon	36	0	1	37
<i>Escherichia coli</i> K12	Bacterium	86	1	1	88
<i>Mycobacterium leprae</i>	Bacterium	45	0	0	45

Ribosomal RNA

Ribosomal RNA molecules form **structural and functional components of ribosomes**, the subcellular **units responsible for protein synthesis**. rRNA constitutes approximately 80–85% of the total RNA in a cell.

rRNA derives from a multicopy ribosomal DNA (rDNA) gene family. In humans these families are localized to the **p arms (i.e., short arms)** of the five acrocentric chromosomes (13, 14, 15, 21, and 22).

CLASSIFICATION OF CHROMOSOMES BASED ON THE POSITION OF CENTROMERE



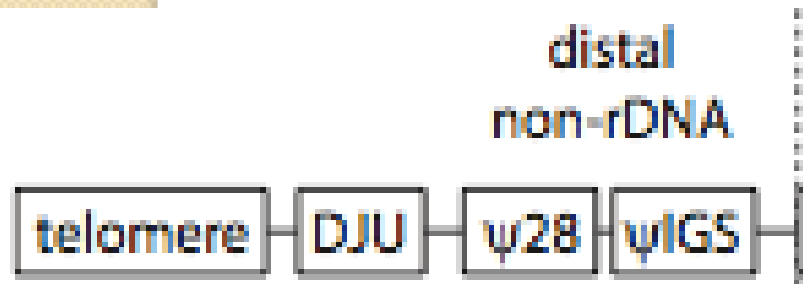
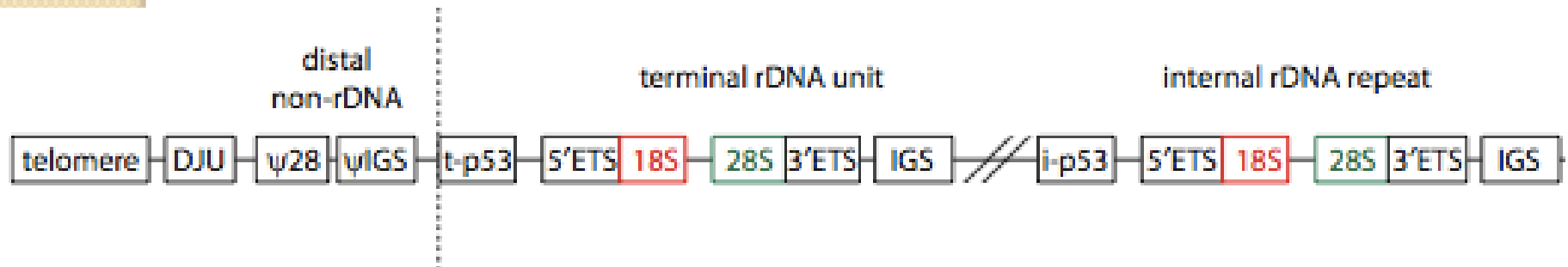
Major forms of rRNA in bacteria and eukaryotes

TABLE 10.3 Major forms of rRNA in bacteria and eukaryotes. S: sedimentation coefficient; MW: molecular weight. Accession numbers are provided for *E. coli* and human rRNAs. Adapted from NCBI and Dayhoff *et al.* (1972, p. D352).

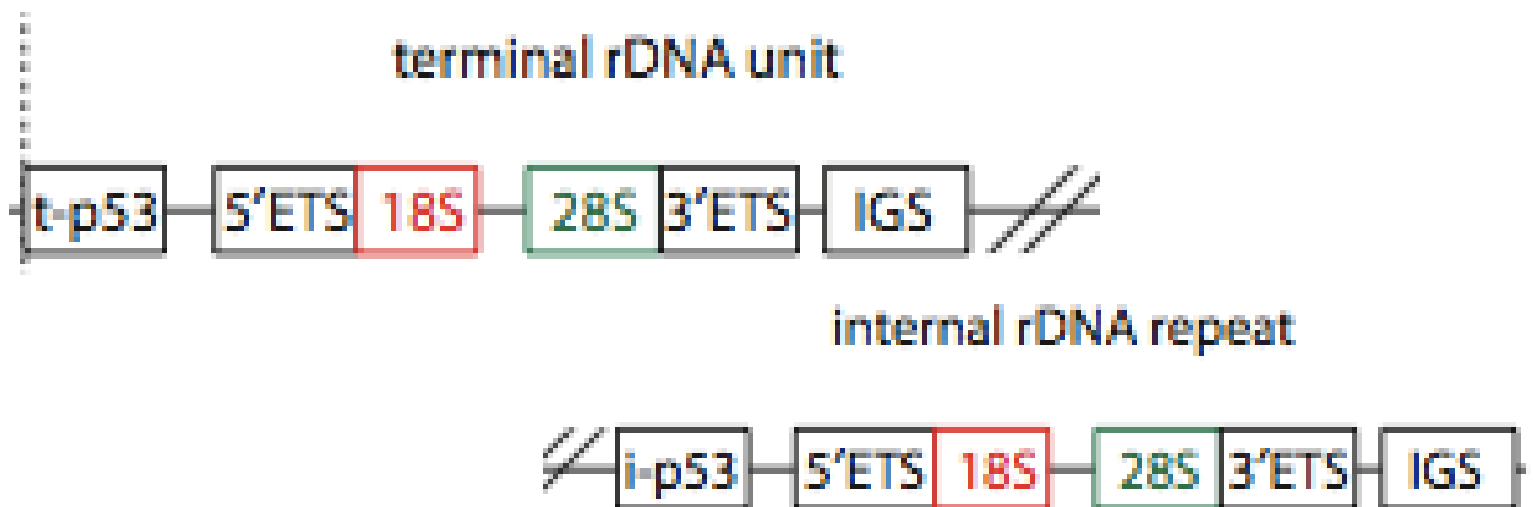
Domain	RN	MW	Ribosomal subunits	rRNA species	Function	Accession number	No. base pairs	RFAM accession
Bacteria	70S	2.6×10^6	30S (small)	16S	Binding mRNA	M25588.1	1504	RF00177
			50S (large)	23S	Peptide bond formation	M25458.1	542	RF02541
				5S		M24300.1	120	RF00001
Eukaryotes	80S	4.3×10^6	40S (small)	18S	Binding mRNA	NR_003286.2	1869	RF01960
			60S (large)	28S	Peptide bond formation	NR_003287.2	5070	RF02543
				5.8S		NR_003285.2	156	RF00002
				5S		NR_023363.1	121	RF00001

Note that the 16S (bacterial) and 18S (eukaryotic) small subunits are commonly used for phylogenetic analyses.

Structure of a eukaryotic ribosomal DNA repeat unit



See GenBank accession U67616. In humans, these rDNA arrays occur on acrocentric chromosome short arms.



Small nuclear RNA (snRNA)

Small nuclear RNA (snRNA) is localized to the nucleus and consists of a family of RNAs that are responsible for **functions such as RNA splicing** (in which **introns are removed from genomic DNA to generate mature mRNA transcripts**) and the **maintenance of telomeres** (chromosome ends).

Examples of human noncoding spliceosomal RNAs

Name	Accession	Chromosome	Length (base pairs)
RNU2-1	NR_002716.3	17 q12-q21	188
RNU4-1	NR_003925.1	12q24.31	144
RNU5F-1	NR_002753.5	1p34.1	116
RNU6-2	NR_002752.2	10p13	107

Small nucleolar RNA (snoRNA) resources

Database	Focus	URL
Plant snoRNA database	<i>Arabidopsis</i> snoRNAs	http://bioinf.scri.sari.ac.uk/cgi-bin/plant_snorna/home
Yeast snoRNA database	H/ACA and C/D box snoRNAs	http://people.biochem.umass.edu/fournierlab/snornadb/main.php
SnoRNABase	human H/ACA and C/D box snoRNAs	https://www-snorna.biotoul.fr//

Database	Focus
Plant snoRNA database	<i>Arabidopsis</i> snoRNAs
Yeast snoRNA database	H/ACA and C/D box snoRNAs
SnoRNABase	human H/ACA and C/D box snoRNAs

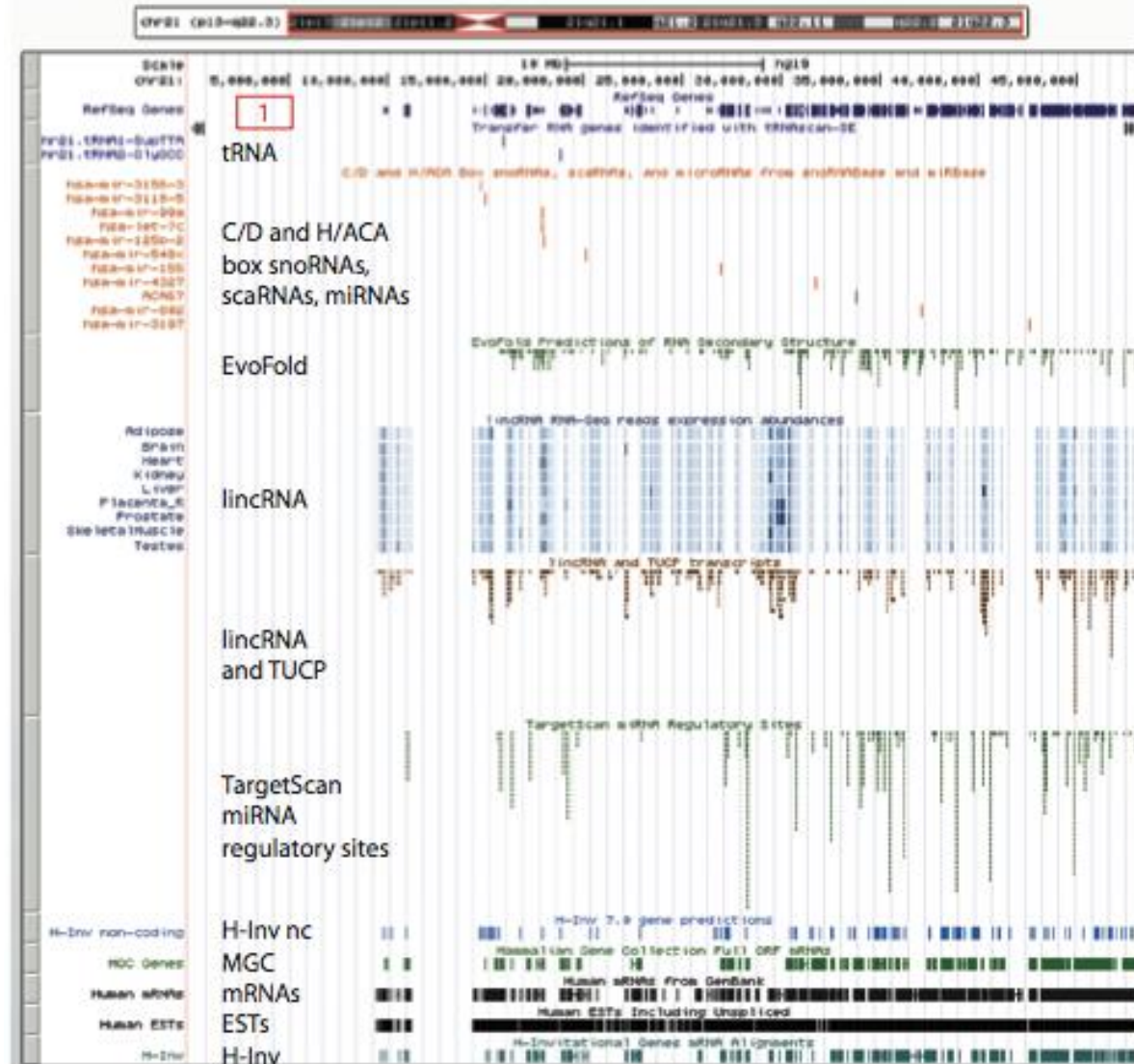
In eukaryotes, ribosome biogenesis occurs in the nucleolus. This process is facilitated by small nucleolar RNAs (snoRNAs), a group of noncoding RNAs that process and modify rRNA and small nuclear spliceosomal RNAs.

Viewing the genomic landscape of noncoding RNAs on human chromosome 2

UCSC Genome Browser on Human Feb. 2009 (GRCh37/hg19) Assembly

move <<< << < > >> >>> zoom in 1.5x 3x 10x base zoom out 1.5x 3x 10x

chr21:1-48,129,895 48,129,895 bp. enter position, gene symbol or search terms go



Outline

Introduction to RNA

Noncoding RNA

Rfam; tRNA; ribosomal RNA; small nuclear RNA; small nucleolar RNA; microRNA; short interfering RNA; long noncoding RNA; UCSC

Introduction to messenger RNA

mRNA; low- and high-throughput technologies; cDNA libraries; full-length cDNA; BodyMap2, GTEx

Microarrays and RNA-seq

Stage 1: experimental design

Stage 2: RNA preparation and probe preparation

Stage 3: data acquisition

Stage 4: data analysis

Stage 5: biological confirmation

Microarray and RNA-seq Databases

Interpretation of RNA analyses

Relationship between DNA, mRNA, and protein

Pervasive nature of transcription

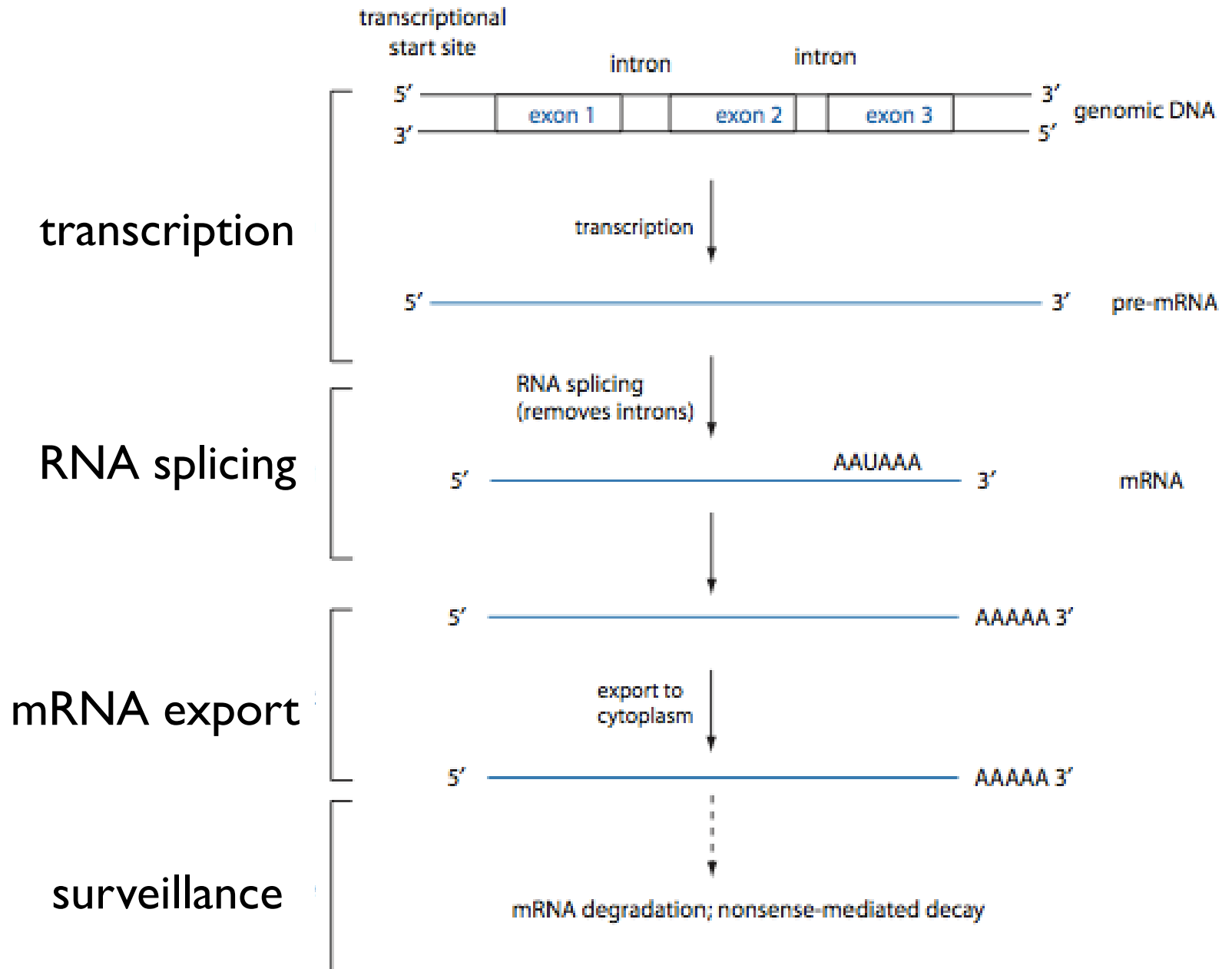
eQtls: genetic basis of variation in gene expression

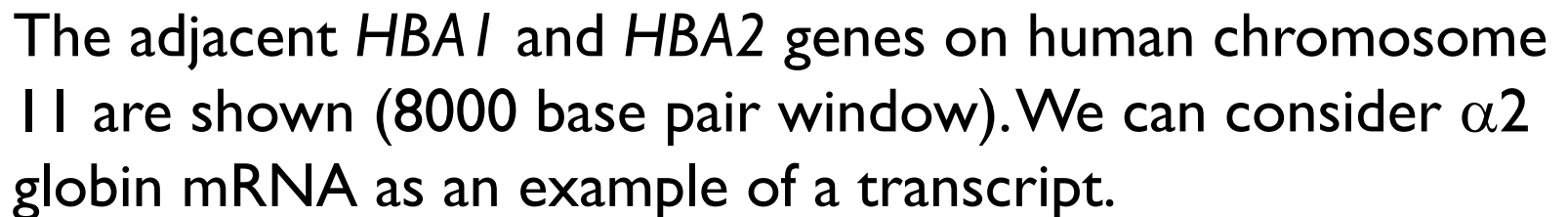
Perspective

Gene expression is context-dependent, and is regulated in several basic ways

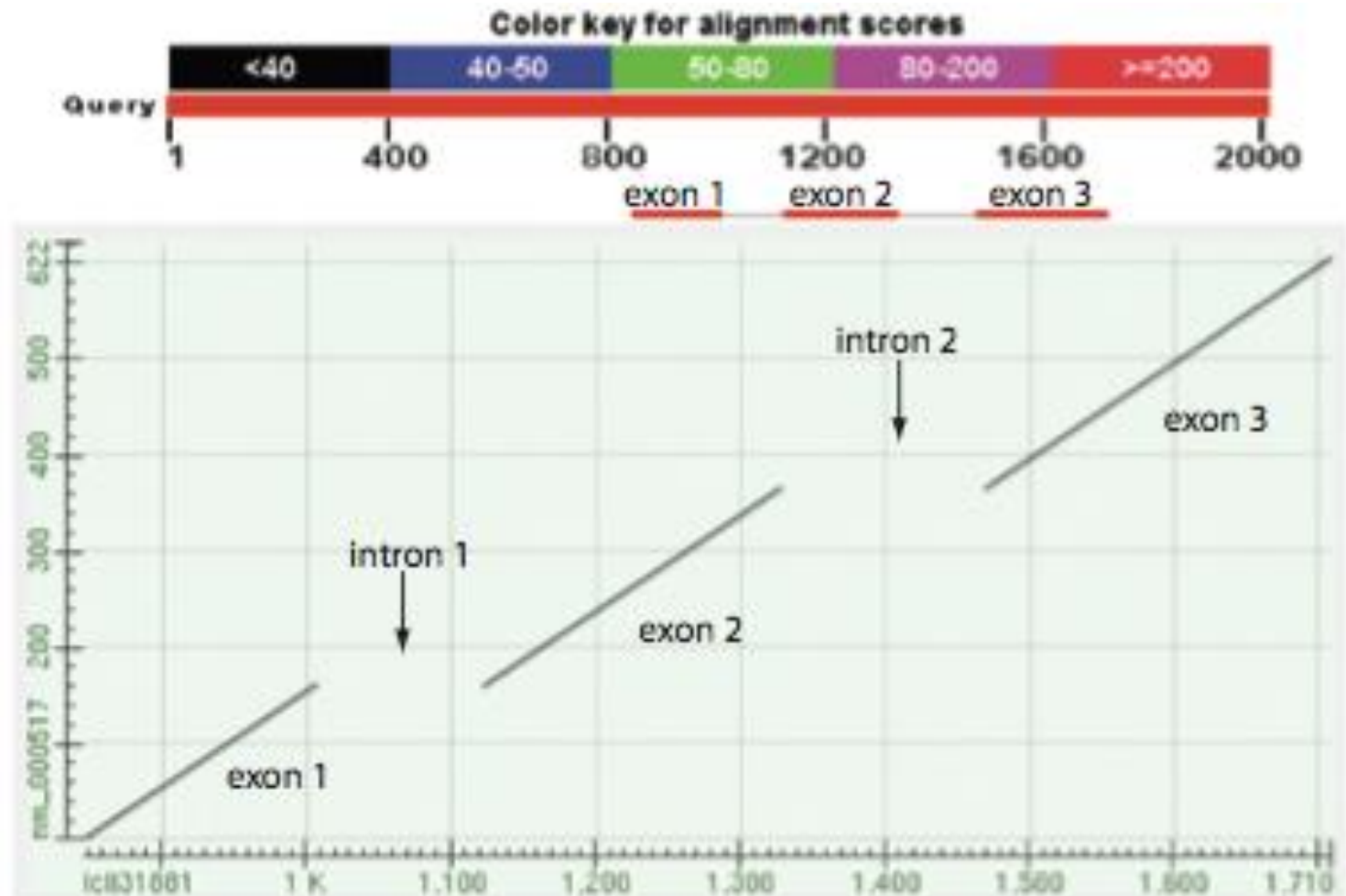
- by **region** (e.g. brain versus kidney)
- in **development** (e.g. fetal versus adult tissue)
- in **dynamic response** to environmental signals
(e.g. immediate-early response genes)
- in disease states
- by gene activity

RNA processing of eukaryotic genes



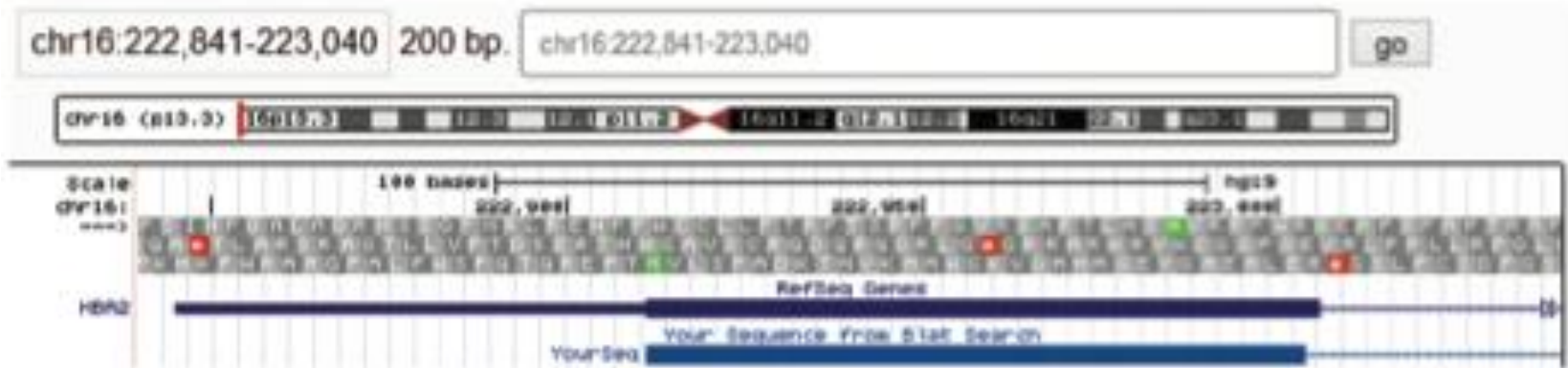


HBA2 mRNA in the context of the corresponding genomic DNA



MegaBLAST of *HBA2* coding sequence (NM_000517.4) to genomic DNA (NT_010393.16, nucleotides 162,000–164,000) reveals positions of exons and introns.

HBA2 mRNA in the context of the corresponding genomic DNA



Exon I of *HBA2* (including nucleotides encoding protein amino terminus).

Low- and high-throughput technologies to study mRNAs

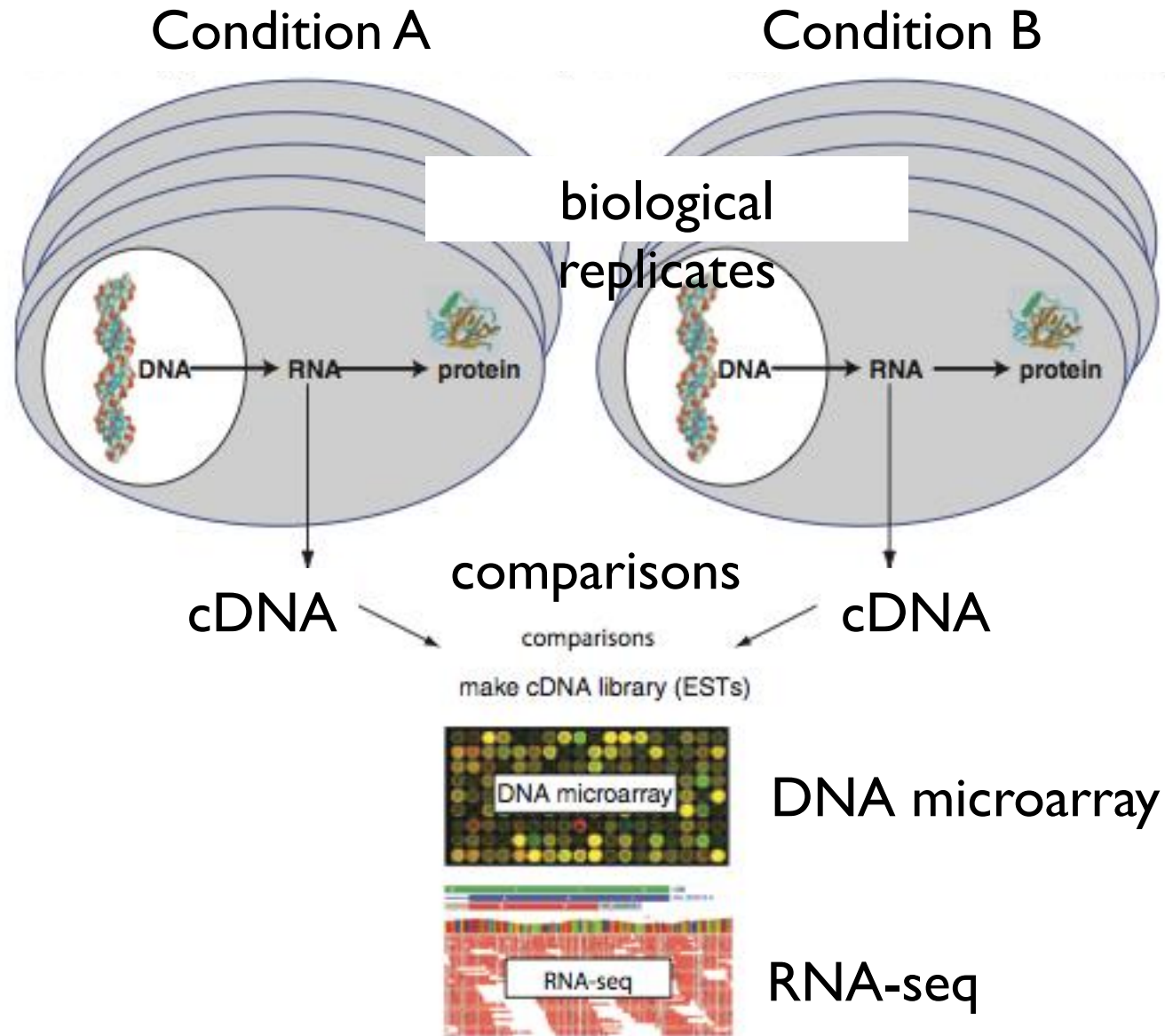
Three techniques for the study of mRNA:

- complementary DNA (cDNA) libraries
- microarrays (e.g. using the Affymetrix platform)
- RNA-seq

Low throughput techniques (Northern blots, PCR) may seem laborious and able to provide only limited amounts of information.

Yet they also serve as trusted “gold standards” and provide crucial validation of high throughput techniques.

Gene expression measured with high-throughput technologies



Analysis of gene expression in cDNA libraries

- The sequencing of cDNA libraries allows the location and quantity of RNA transcripts to be measured.
- cDNA inserts, called expressed sequence tags (ESTs), are sequenced.
- The UniGene database partitions ESTs into nonredundant clusters that generally correspond to expressed genes.
- Each cluster has some number of sequences associated with it, from one (*singletons*) to ~50,000

Cluster sizes for human entries in UniGene

Cluster size	Number of clusters	Example(s) of genes in cluster
1	64,371	
2	12,760	
3-4	10,859	Transcribed locus, strongly similar to NP_032247.1 hemoglobin subunit epsilon-Y2 [Mus musculus]
5-8	10,637	Transcribed locus, strongly similar to NP_001077424.1 hemoglobin alpha, adult chain 2 [Mus musculus]
9-16	7,177	Hemoglobin, theta 1; hemoglobin, beta pseudogene 1
17-32	4,815	Hemoglobin, mu; neuroglobin
33-64	4,557	Hemoglobin, zeta
65-128	4,117	Hemoglobin, delta
129-256	3,889	Hemoglobin, epsilon 1; cytoglobin
257-512	3,858	
513-1024	1,982	
1,025-2,048	729	Hemoglobin, alpha 1; myoglobin; hemoglobin, gamma A
2,049-4,096	224	Hemoglobin, beta; hemoglobin, gamma G
4,097-8,192	56	Hemoglobin, alpha 2
8,193-16,384	20	Albumin, GAPDH; ubiquitin C; tubulin, alpha 1b; ferritin, light polypeptide
16,385-32,768	4	Actin, beta; myelin basic protein; Eukaryotic translation elongation factor 1 alpha 1; Uncharacterized LOC100507412
32,769-65,536	1	EEF1A1

- About 64,000 clusters have just one EST (in build 236, *Homo sapiens*)
- Just one cluster has >32,000 ESTs (very highly expressed)

Ten largest cluster sizes in UniGene for human entries

UniGene Identifier	Cluster size	Gene symbol	Gene name
Hs.586423	48,000	<i>EEF1A1</i>	Eukaryotic translation elongation factor 1 alpha 1
Hs.535192	27,000	<i>EEF1A1</i>	Eukaryotic translation elongation factor 1 alpha 1
Hs.520640	26,000	<i>ACTB</i>	Actin, beta
Hs.551713	21,000	<i>MBP</i>	Myelin basic protein
Hs.426704	20,000	<i>LOC100507412</i>	Uncharacterized LOC100507412
Hs.520348	16,000	<i>UBC</i>	Ubiquitin C
Hs.418167	16,000	<i>ALB</i>	Albumin
Hs.524390	16,000	<i>TUBA1B</i>	Tubulin, alpha 1b
Hs.510635	16,000	<i>IGHG1</i>	Immunoglobulin heavy constant gamma 1 (G1m marker)
Hs.544577	15,000	<i>GAPDH</i>	Glyceraldehyde-3-phosphate dehydrogenase
Hs.180414	15,000	<i>HSPA8</i>	Heat shock 70kDa protein 8
Hs.370247	15,000	<i>APLP2</i>	Amyloid beta (A4) precursor-like protein 2

Ten largest cluster sizes in UniGene for nonhuman entries

UniGene Identifier	Species	Cluster size	Gene Name
Cin.19067	<i>Ciona intestinalis</i> (vase tunicate; yellow sea squirt)	48,000	Clone:citb001e24, full insert sequence
Bfl.2106	<i>Branchiostoma floridae</i> (Florida lancelet)	31,000	Transcribed locus, strongly similar to NP_007768.1 NADH dehydrogenase subunit 1
Bt.107724	<i>Bos taurus</i> (cow)	22,000	Chymotrypsinogen B1-like
At.46639	<i>Arabidopsis thaliana</i> (thale cress)	16,000	Ribulose biphosphate carboxylase small chain 1A
Cin.30513	<i>Ciona intestinalis</i>	15,000	ATP-binding cassette sub-family D member 2-like
Dr.31797	<i>Danio rerio</i> (zebrafish)	13,000	Eukaryotic translation elongation factor 1 alpha 1, like 1
Dr.75552	<i>Danio rerio</i>	13,000	Actin, alpha, cardiac muscle 1b
Rn.202968	<i>Rattus norvegicus</i> (Norway rat)	13,000	Albumin
Ta.11048	<i>Triticum aestivum</i> (bread wheat)	13,000	Small subunit
Ssc.6512	<i>Sus scrofa</i> (pig)	12,000	Mitochondrial ATPase 6 mRNA, L transcript, partial sequence

- Search query `||700:65536[sequence count] NOT txid9606[organism]`
- These represent highly expressed transcripts

Outline

Introduction to RNA

Noncoding RNA

Rfam; tRNA; ribosomal RNA; small nuclear RNA; small nucleolar RNA; microRNA; short interfering RNA; long noncoding RNA; UCSC

Introduction to messenger RNA

mRNA; low- and high-throughput technologies; cDNA libraries; full-length cDNA; BodyMap2, GTEx

Microarrays and RNA-seq

Stage 1: experimental design

Stage 2: RNA preparation and probe preparation

Stage 3: data acquisition

Stage 4: data analysis

Stage 5: biological confirmation

Microarray and RNA-seq Databases

Interpretation of RNA analyses

Relationship between DNA, mRNA, and protein

Pervasive nature of transcription

eQtls: genetic basis of variation in gene expression

Perspective

Full-length cDNAs and expression patterns

- There are many resources to purchase (or study) full-length cDNAs (e.g. FANTOM, Mammalian Gene Collection)(accessed via NCBI Gene)

Measuring gene expression across the body

- The Genotype-Tissue Expression (GTEx) project catalogs tissue-specific gene expression across the human body.
- BodyMap2 measures gene expression across 16 tissues using RNA-seq

Outline

Introduction to RNA

Noncoding RNA

Rfam; tRNA; ribosomal RNA; small nuclear RNA; small nucleolar RNA; microRNA; short interfering RNA; long noncoding RNA; UCSC

Introduction to messenger RNA

mRNA; low- and high-throughput technologies; cDNA libraries; full-length cDNA; BodyMap2, GTEx

Microarrays and RNA-seq

Stage 1: experimental design

Stage 2: RNA preparation and probe preparation

Stage 3: data acquisition

Stage 4: data analysis

Stage 5: biological confirmation

Microarray and RNA-seq Databases

Interpretation of RNA analyses

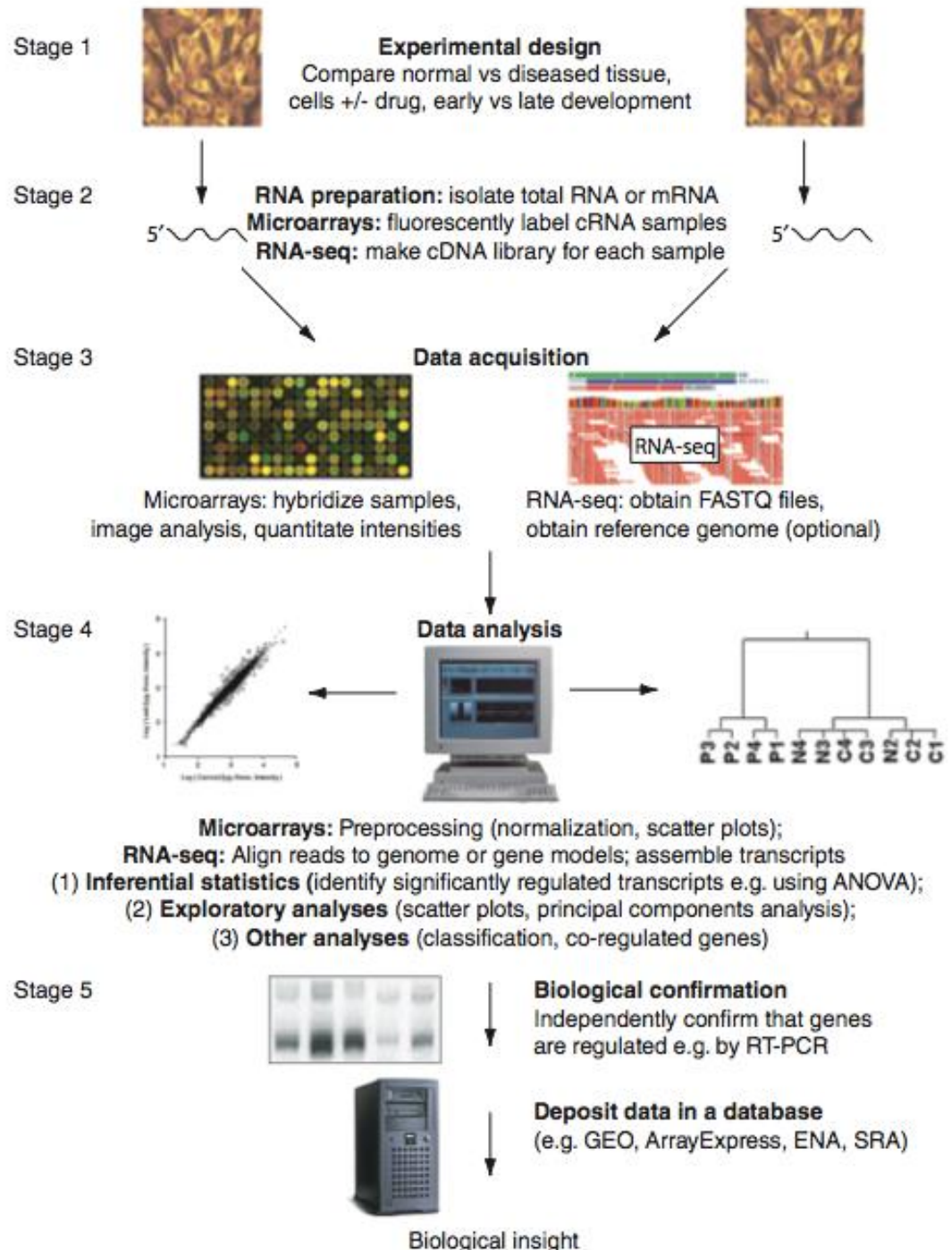
Relationship between DNA, mRNA, and protein

Pervasive nature of transcription

eQtls: genetic basis of variation in gene expression

Perspective

Overview of the process of generating high-throughput gene expression data using microarrays or RNA-seq



Outline

Introduction to RNA

Noncoding RNA

Rfam; tRNA; ribosomal RNA; small nuclear RNA; small nucleolar RNA; microRNA; short interfering RNA; long noncoding RNA; UCSC

Introduction to messenger RNA

mRNA; low- and high-throughput technologies; cDNA libraries; full-length cDNA; BodyMap2, GTEx

Microarrays and RNA-seq

Stage 1: experimental design

Stage 2: RNA preparation and probe preparation

Stage 3: data acquisition

Stage 4: data analysis

Stage 5: biological confirmation

Microarray and RNA-seq Databases

Interpretation of RNA analyses

Relationship between DNA, mRNA, and protein

Pervasive nature of transcription

eQtls: genetic basis of variation in gene expression

Perspective

Stage 1: Experimental design

Stage 2: RNA preparation

Stage 3: Hybridization to DNA arrays

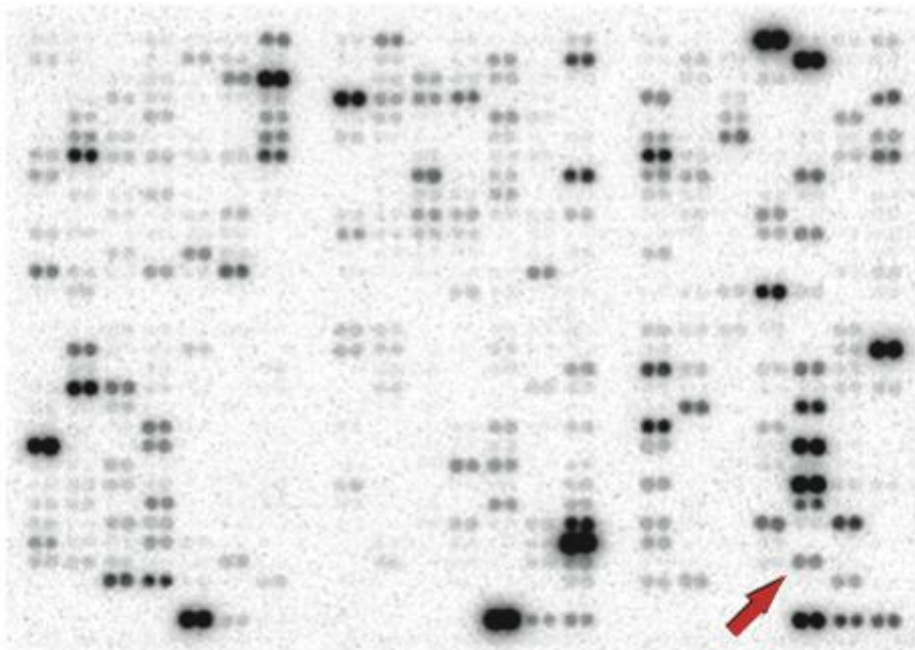
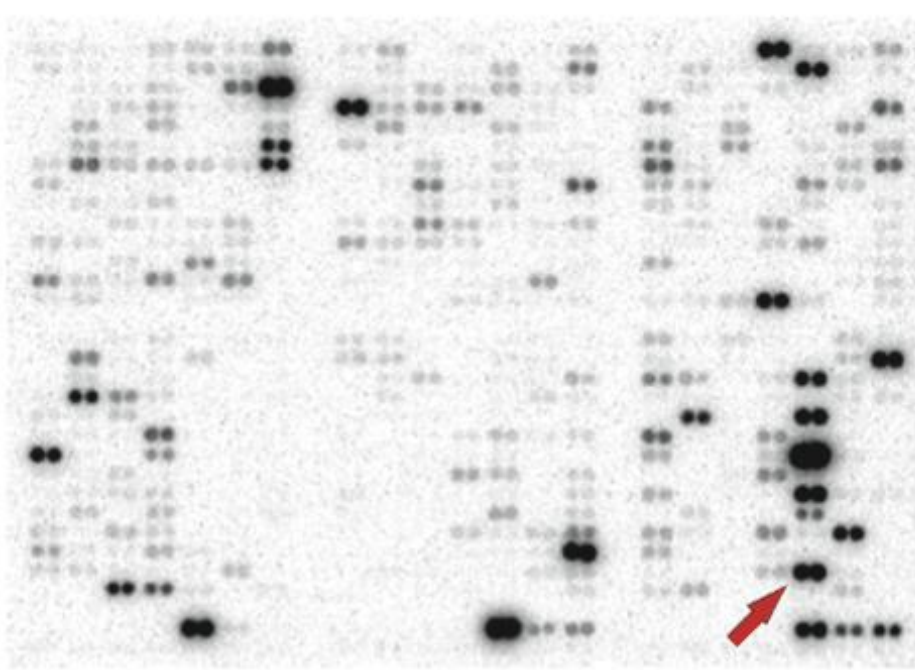
[1] Be sure to use enough biological replicates (typically $n \geq 3$ per group). Consult a statistician if you're unsure.

[2] RNA extraction, conversion, labeling, hybridization: evaluate and avoid systematic artifacts (avoid batch effects). Be sure to create an appropriately balanced, randomized experimental design.

[3] Microarrays typically consist of oligonucleotides (deposited by photolithography), and samples are cRNA or cDNA with fluorescent tags.

Microarray experiment with radioactive probes

Technology using radioactive probes is now obsolete, but this figure illustrates the essential nature of differential gene expression in two samples.



Arrow indicates a transcript expressed at a lower level (in the bottom panel)

Outline

Introduction to RNA

Noncoding RNA

Rfam; tRNA; ribosomal RNA; small nuclear RNA; small nucleolar RNA; microRNA; short interfering RNA; long noncoding RNA; UCSC

Introduction to messenger RNA

mRNA; low- and high-throughput technologies; cDNA libraries; full-length cDNA; BodyMap2, GTEx

Microarrays and RNA-seq

Stage 1: experimental design

Stage 2: RNA preparation and probe preparation

Stage 3: data acquisition

Stage 4: data analysis

Stage 5: biological confirmation

Microarray and RNA-seq Databases

Interpretation of RNA analyses

Relationship between DNA, mRNA, and protein

Pervasive nature of transcription

eQtls: genetic basis of variation in gene expression

Perspective

Stage 4: Microarray data analysis

Hypothesis testing

How can arrays be compared?

- Which RNA transcripts (genes) are regulated?
- Are differences authentic?
- What are the criteria for statistical significance?

Clustering

- Are there meaningful patterns in the data (e.g. groups)?

Classification

- Do RNA transcripts predict predefined groups, such as disease subtypes?

Stage 5: Biological confirmation

Microarray experiments can be thought of as “hypothesis-generating” experiments.

The differential up- or down-regulation of specific RNA transcripts can be measured using independent assays such as

- Northern blots
- polymerase chain reaction (RT-PCR)
- in situ hybridization


Stage 6: Microarray databases

There are two main repositories: Gene expression Omnibus (GEO) at NCBI and ArrayExpress at the European Bioinformatics Institute (EBI).


Minimum Information About a Microarray Experiment (MIAME) guidelines are followed to describe experiments:

- ▶ experimental design
- ▶ microarray design
- ▶ sample preparation
- ▶ hybridization procedures
- ▶ image analysis
- ▶ controls for normalization


Entrez provides access to GEO Profiles and Datasets


**NCBI** Resources ▾ How To ▾

pevsner My NCBI Sign Out

**NCBI**
National Center for
Biotechnology Information

All Databases ▾ globin






**NCBI**















 *Entrez, The Life Sciences Search Engine.*

HOME SEARCH SITE MAP PubMed All Databases Human Genome GenBank Map Viewer BLAST

Search across databases [Help](#)

- Result counts displayed in gray indicate one or more terms not found

111344  PubMed: biomedical literature citations and abstracts	379  Books: online books
28150  PubMed Central: free, full text journal articles	151  OMIM: online Mendelian Inheritance in Man
4  Site Search: NCBI web and FTP sites	

9493  Nucleotide: Core subset of nucleotide sequence records	none  dbGaP: genotype and phenotype
3172  EST: Expressed Sequence Tag records	223  UniGene: gene-oriented clusters of transcript sequences
3  GSS: Genome Survey Sequence records	13  CDD: conserved protein domain database
21158  Protein: sequence database	105  UniSTS: markers and mapping data
317  Genome: whole genome sequences	102  PopSet: population study data sets
870  Structure: three-dimensional macromolecular structures	3129  GEO Profiles: expression and molecular abundance profiles
none  Taxonomy: organisms in GenBank	353  GEO DataSets: experimental sets of GEO data

GEO profiles: how was my favorite RNA transcript expressed across thousands of experiments?

NCBI Resources ☒ How To ☒ pevsnr My NCBI Sign Out

GEO Profiles

[Save search](#) [Limits](#) [Advanced](#)

[Display Settings:](#) ☒ Summary, 20 per page, Sorted by Subgroup effect

[Send to:](#) ☒

[Filters:](#) [Manage Filters](#)

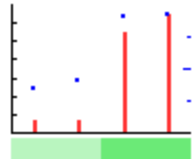
Results: 1 to 20 of 3129

<< First < Prev Page **1** of 157 [Next >](#) [Last >>](#)

☐ **1:** [GDS1581 record](#) | [GPL86 rc_AI012182_s_at](#)
[[Rattus norvegicus](#)]

4 samples [Profile Neighbors](#), [Chromosome Neighbors](#),
[Homologs](#), [Links](#)

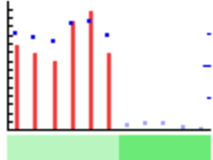
Annotation: [LOC100134871](#): beta globin minor gene rCG_39884 (multiple annotations exist)
Reporter: [AI012182](#)
Experiment: Embryonic kidney: ureteric bud and metanephric mesenchyme (RG-U34B), Expression profiling by array, count



☐ **2:** [GDS954 record](#) | [GPL341 1371245_a_at](#)
[[Rattus norvegicus](#)]

11 samples [Profile Neighbors](#), [Chromosome Neighbors](#),
[Homologs](#), [Links](#)


Annotation: [LOC100134871](#): beta globin minor gene rCG_39884 (multiple annotations exist)
Reporter: [BI287300](#)
Experiment: Ketogenic diet effect on brain hippocampus, Expression profiling by array, count



☐ **3:** [GDS3189 record](#) | [GPL85 rc_AA860014_i_at](#) [[Rattus norvegicus](#)]

4 samples [Profile Neighbors](#), [Links](#)

Annotation: [UI-R-E0-ca-c-11-0-UI.s1](#) UI-R-E0 Rattus norvegicus cDNA clone UI-R-E0-ca-c-11-0-UI 3-similar to gb|M13125|MUSHBX Mouse alpha-like embryonic x globin chain, mRNA, mRNA sequence



Profile data

[Download profile data](#)

Profile pathways

[Find pathways](#)

Find related data

Database:

[Find items](#)

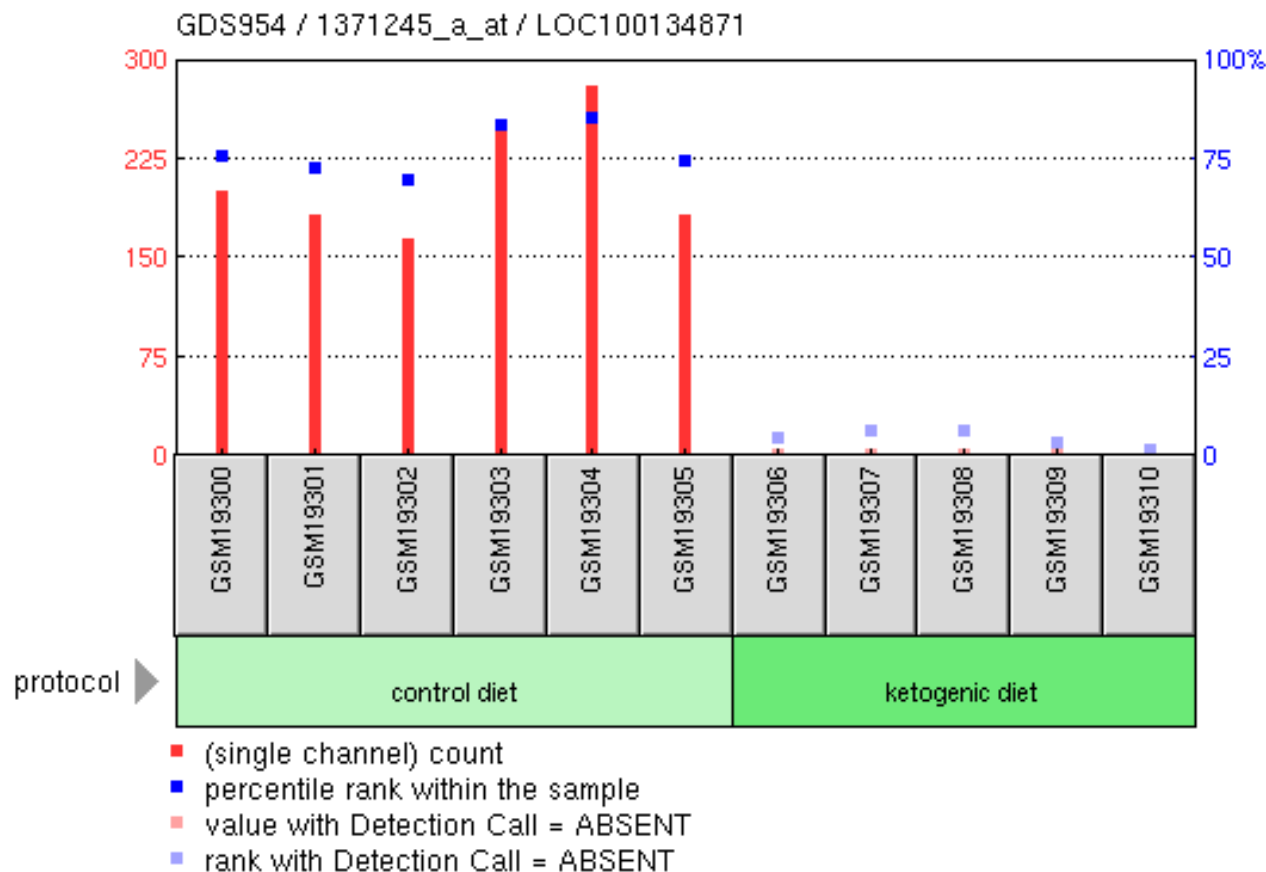
Search details

[globin](#)[All Fields]

GEO profiles: example of a globin RNA transcript expressed at low levels in an experimental condition

Title: [GDS954](#) / 1371245_a_at / LOC100134871 / *Rattus norvegicus*

Summary: Expression profiling of brain hippocampi of animals fed the ketogenic diet (KD). KD is an anticonvulsant treatment used to manage medically intractable epilepsies. Results provide insight into the anticonvulsant action of KD.



Outline

Introduction to RNA

Noncoding RNA

Rfam; tRNA; ribosomal RNA; small nuclear RNA; small nucleolar RNA; microRNA; short interfering RNA; long noncoding RNA; UCSC

Introduction to messenger RNA

mRNA; low- and high-throughput technologies; cDNA libraries; full-length cDNA; BodyMap2, GTEx

Microarrays and RNA-seq

Stage 1: experimental design

Stage 2: RNA preparation and probe preparation

Stage 3: data acquisition

Stage 4: data analysis

Stage 5: biological confirmation

Microarray and RNA-seq Databases

Interpretation of RNA analyses

Relationship between DNA, mRNA, and protein

Pervasive nature of transcription

eQTLs: genetic basis of variation in gene expression

Perspective

The relationship between DNA, mRNA, and protein levels

Once mRNA levels are present at elevated or reduced levels, are the corresponding proteins differentially expressed in a similar manner? Perhaps surprisingly, there appears to be only a weak positive correlation between mRNA and protein levels.

Several groups have reported described correlation coefficients that were relatively high when highly abundant proteins were considered (e.g., $r = 0.935$, $r = 0.86$ in two studies) but lower when highly abundant proteins were excluded (e.g., $r = 0.36$, $r = 0.49$, $r = 0.21$, $r = 0.18$).

Weak correlations could be due to RNA structural effects, regulatory noncoding RNAs, codon bias, variable protein half-lives, and experimental error.

The pervasive nature of transcription

Strong evidence for pervasive transcription comes from the ENCODE project (ENCODE Consortium, 2007; Djebali *et al.*, 2012). Transcriptional activity was measured using a series of technologies.

Conclusions include the following:

- 62.1% and 74.7% of the human genome is spanned by processed or primary transcripts, respectively;
- genes express 10–12 isoforms per cell line;
- coding RNA transcripts tend to be cytosolic, while noncoding transcripts are localized to the nucleus;
- ~6% of annotated coding and noncoding transcripts overlap small noncoding RNAs.

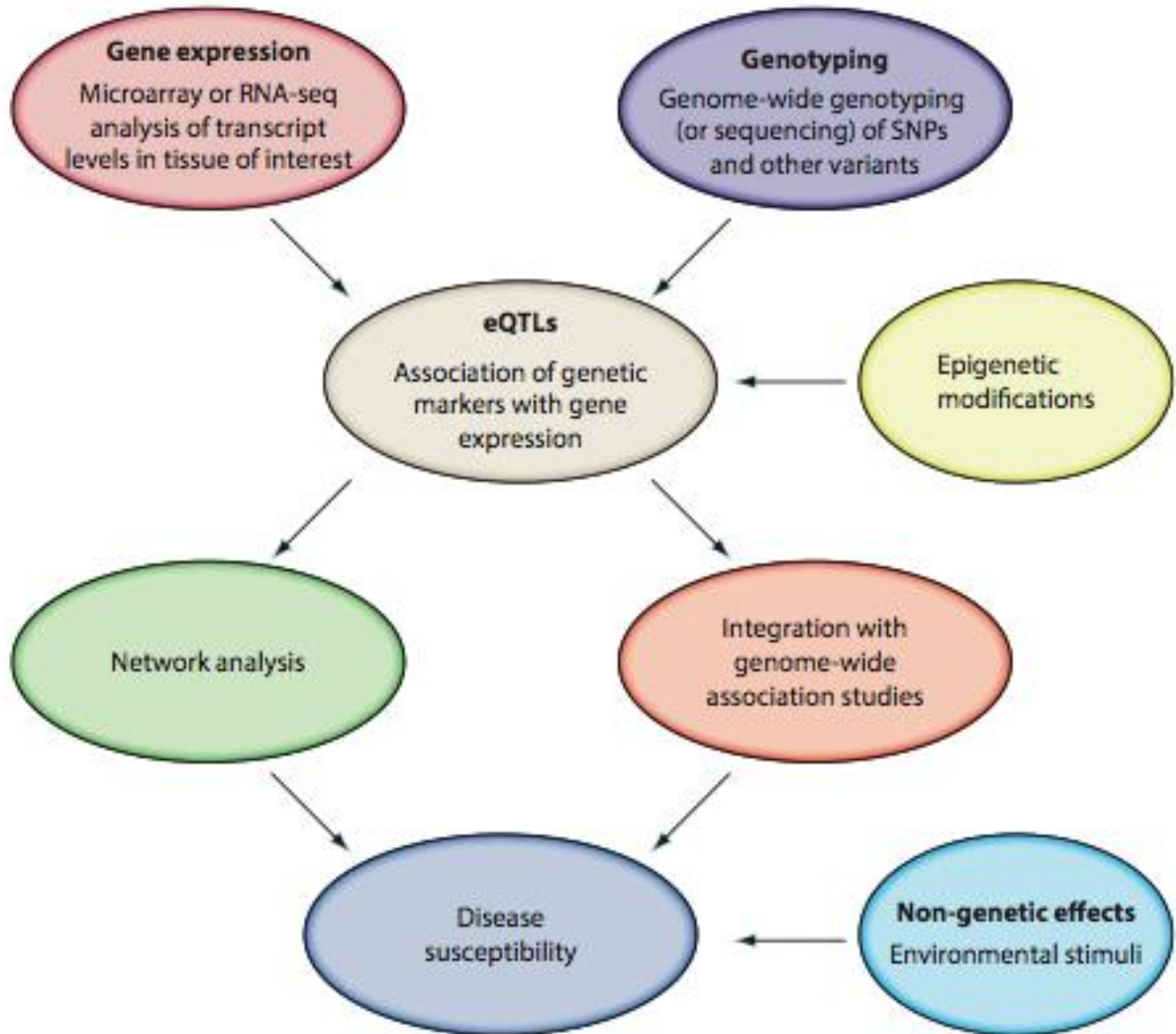
Much of the genome is transcribed. Some of this transcription is certain to be biologically relevant, while in other cases it is likely to represent biological “noise” associated with low levels of transcription.

eQTLs: expression quantitative trait loci

mRNA expression is a quantitative trait that can be described for a given cell type and physiological state in an organism. Furthermore, variants in genomic DNA may impact mRNA expression. **Expression quantitative trait loci (eQTLs)** are genomic loci that control expression levels.

Two main types of **control regions** have been found: (1) **cis-eQTLs** are genomic loci that influence the expression of transcripts expressed from neighboring genes **within some distance (such as 1 Mb or less)**, and may undergo allele-specific expression; and (2) **Trans-eQTLs** act on transcripts expressed from genes that are **farther away or on another chromosome**. eQTLs could affect transcription directly or indirectly, for example by **altering the sequence of a transcription factor binding site** that controls a gene's expression proximally or distally.

Expression quantitative trait loci (eQTLs)



Outline

Introduction to RNA

Noncoding RNA

Rfam; tRNA; ribosomal RNA; small nuclear RNA; small nucleolar RNA; microRNA; short interfering RNA; long noncoding RNA; UCSC

Introduction to messenger RNA

mRNA; low- and high-throughput technologies; cDNA libraries; full-length cDNA; BodyMap2, GTEx

Microarrays and RNA-seq

Stage 1: experimental design

Stage 2: RNA preparation and probe preparation

Stage 3: data acquisition

Stage 4: data analysis

Stage 5: biological confirmation

Microarray and RNA-seq Databases

Interpretation of RNA analyses


Relationship between DNA, mRNA, and protein

Pervasive nature of transcription

eQtls: genetic basis of variation in gene expression

Perspective

Perspective



Genes in all organisms are expressed in a variety of developmental, environmental, or physiological conditions. The field of functional genomics includes the high-throughput study of gene expression.

Before the arrival of this new approach, the expression of one gene at a time was typically studied. Functional genomics may reveal the transcriptional program of entire genomes, allowing a global view of cellular function.