CS273A

Lecture 3: Non Coding Genes

MW 1:30-2:50pm in Clark **S361*** (behind Peet's)

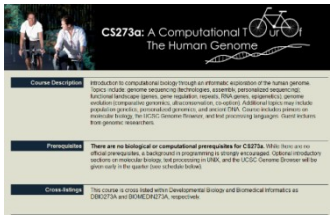
Profs: Serafim Batzoglou & Gill Bejerano

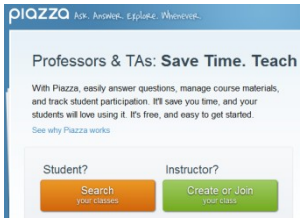
CAs: Karthik Jagadeesh & Johannes Birgmeier

* Mostly: track on website/piazza

Announcements

- http://cs273a.stanford.edu/
  - Lecture slides, problem sets, etc.
- Course communications via Piazza
  - Auditors please sign up too



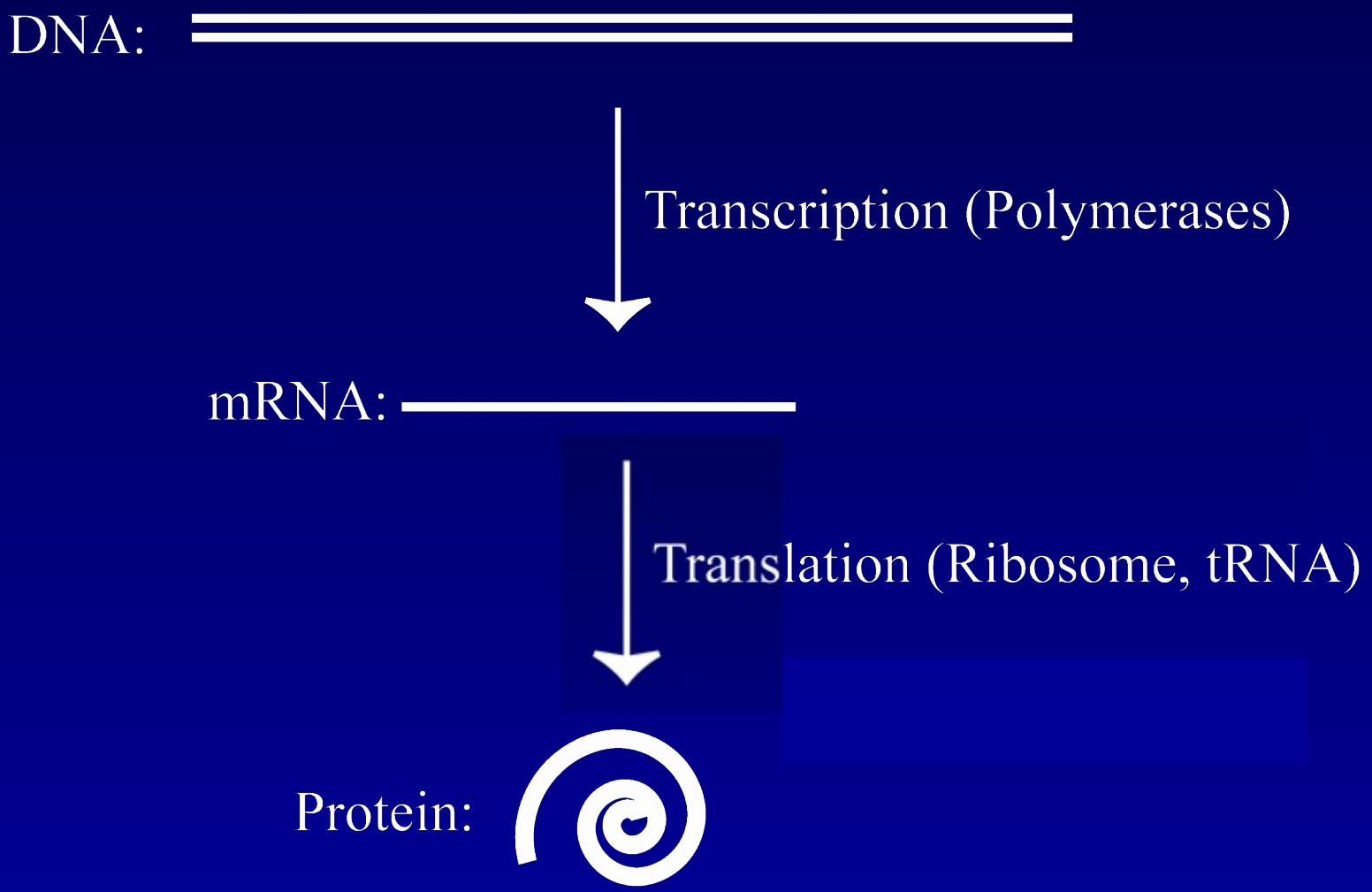


- Second Tutorial this Fri 10/2.
  UCSC genome browser. We recommend bringing your laptop.

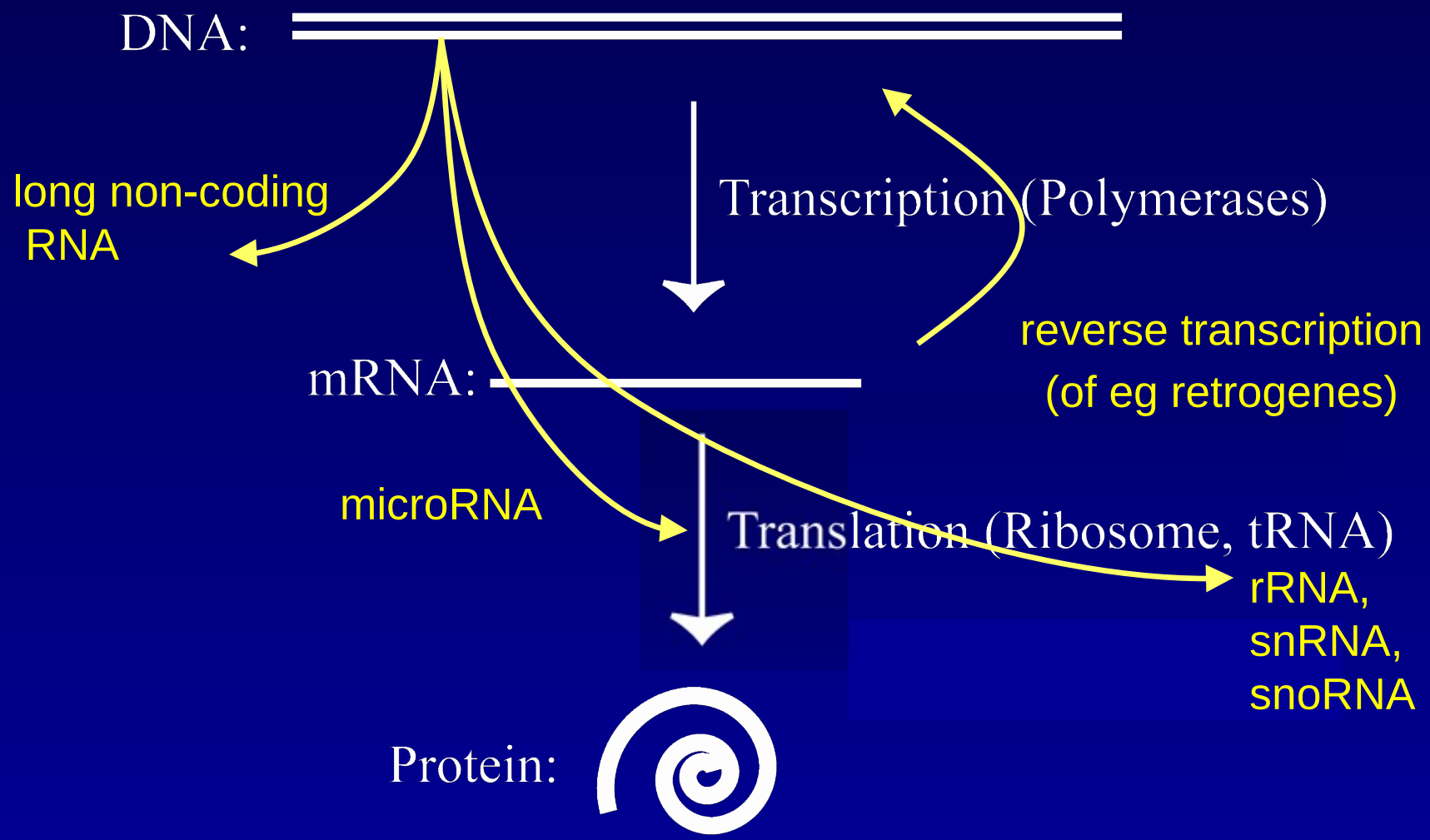
- We are starting to take attendance today.

“non coding” RNAs (ncRNA)

Central Dogma of Biology:



Active forms of “non coding” RNA



What is ncRNA?

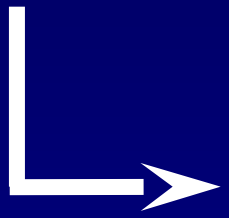
- Non-coding RNA (ncRNA) is an RNA that functions without being translated to a protein.
- Known roles for ncRNAs:
  - RNA catalyzes excision/ligation in introns.
  - RNA catalyzes the maturation of tRNA.
  - RNA catalyzes peptide bond formation.
  - RNA is a required subunit in telomerase.
  - RNA plays roles in immunity and development (RNAi).
  - RNA plays a role in dosage compensation.
  - RNA plays a role in carbon storage.
  - RNA is a major subunit in the SRP, which is important in protein trafficking.
  - RNA guides RNA modification.
- RNA can do so many different functions, it is thought in the beginning there was an RNA World, where RNA was both the information carrier and active molecule.

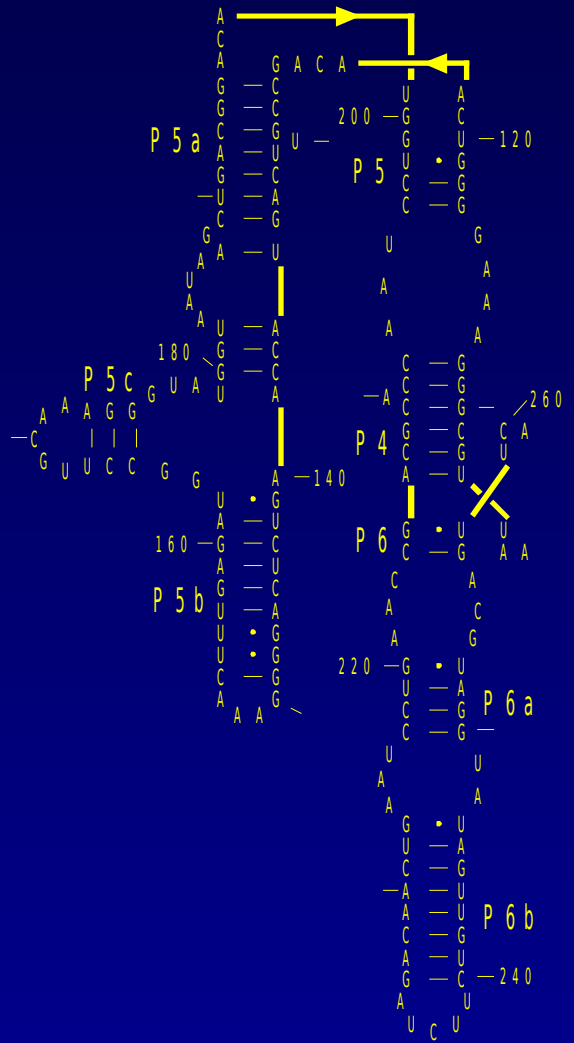
“non coding” RNAs (ncRNA) subtype:

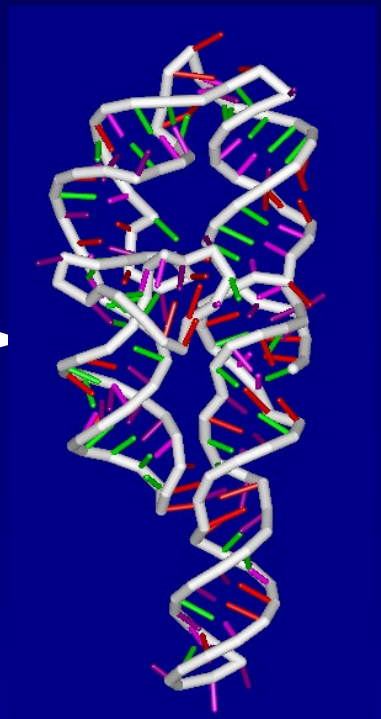
Small structural RNAs (ssRNA)

ssRNA Folds into Secondary and 3D Structures

AAUUGCGGGAAAGGGGUCAA
CAGCCGUUCAGUACCAAGUC
UCAGGGGAAACUUUGAGAUG
GCCUUGCAAAGGGUAUGGUA
AUAAGCUGACGGACAUGGUC
CUAACCACGCAGCCAAGUCC
UAAGUCAACAGAUCUUCUGU
UGAUAUGGAUGCAGUUCA





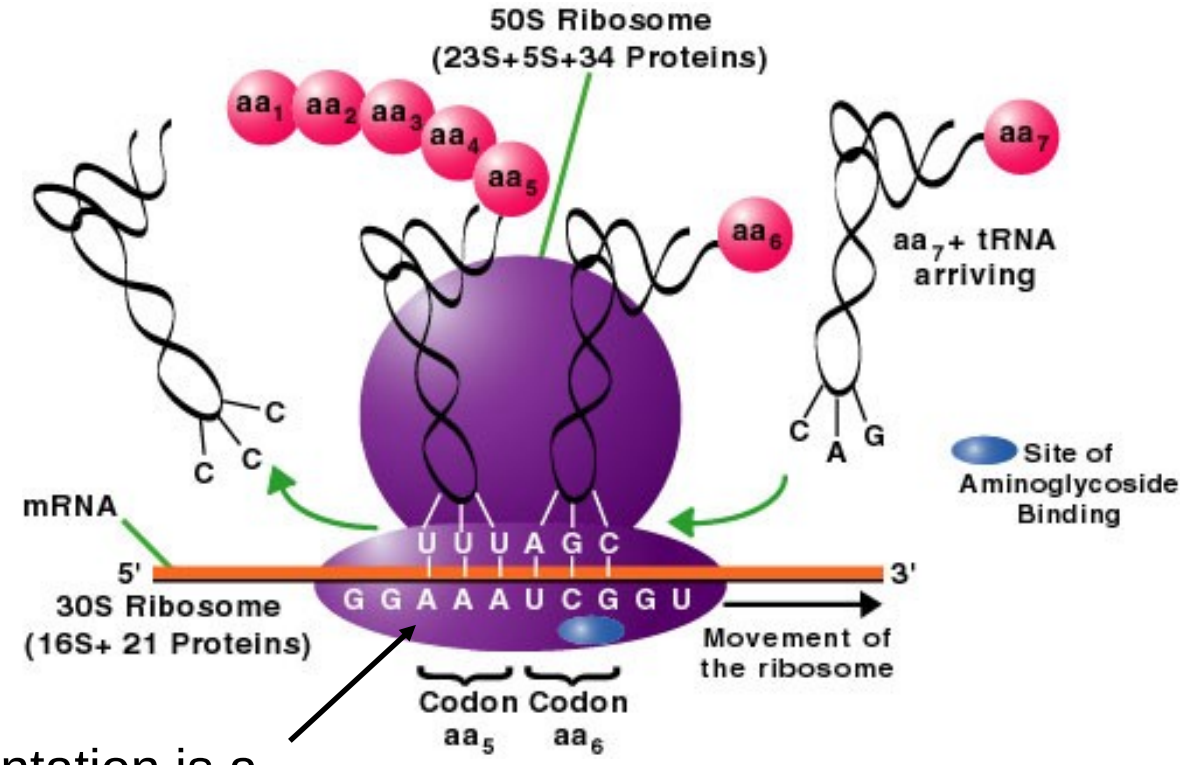


Computational Challenge: predict 2D & 3D structure from sequence (1D).

Waring & Davies. (1984) *Gene 28*: 277.

Cate, *et al.* (Cech & Doudna). (1996) *Science 273*:1678.

For example, tRNA Activity

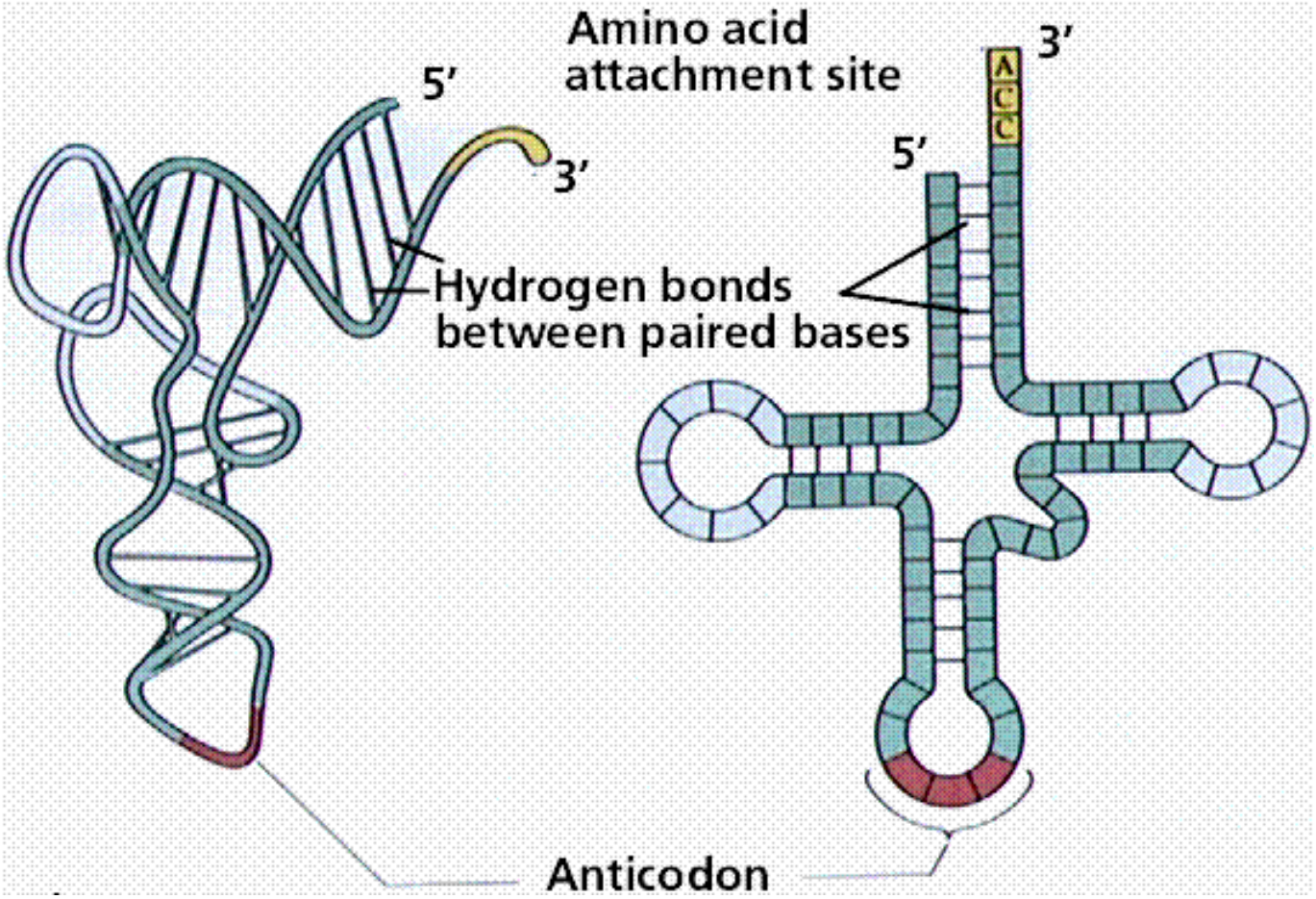


Complementation is a "brilliant" way for the Genome to "get things done".

Replication, Transcription, Translation, …

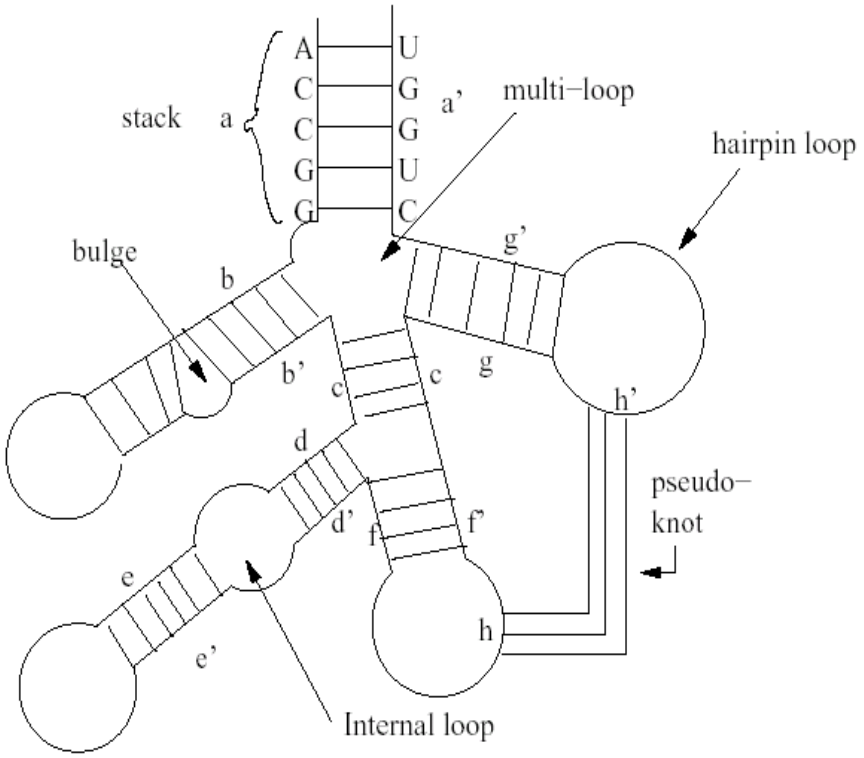
Compliment(Compliment(Identity)) = Identity

tRNA Structure



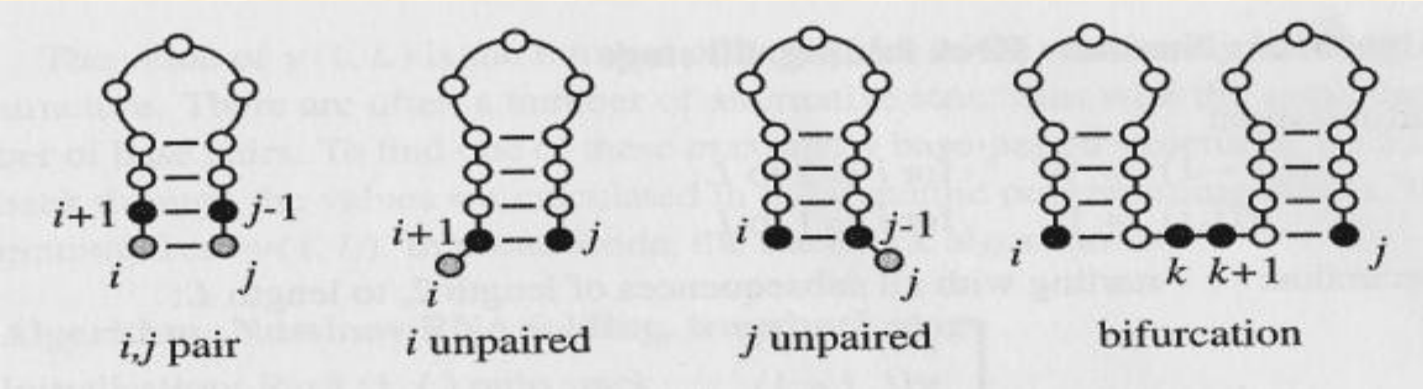
ssRNA structure rules

- Canonical basepairs:
  - Watson-Crick basepairs:
    - G ≡ C
    - A = U
  - Wobble basepair:
    - G – U
- Stacks: continuous nested basepairs. (energetically favorable)
- Non-basepaired loops:
  - Hairpin loop
  - Bulge
  - Internal loop
  - Multiloop
- Pseudo-knots



Ab initio RNA structure prediction: lots of Dynamic Programming

- Objective: Maximizing # of base pairings (Nussinov *et al*, 1978)



$$S(i,j) = max \begin{cases} S(i+1,j-1) + w(i,j) & (1) \\ S(i+1,j) & (2) \\ S(i,j-1) & (3) \\ max_{i<k<j} S(i,k) + S(k+1,j) & (4) \end{cases}$$

simple model:
ω(i, j) = 1 iff GC|AU|GU
fancier model:
GC > AU > GU

Dynamic programming

- Compute S(i,j) recursively (dynamic programming)
  - Compares a sequence against itself in a dynamic programming matrix
- Three steps

1. Initialization

	G	G	G	A	A	A	U	C	C
G	0								
G	0	0							
G		0	0						
A			0	0					
A				0	0				
A					0	0			
U						0	0		
C							0	0	
C								0	0

Example:

GGGAAAUCC

$$S(i,i) = 0 \quad \forall \quad 1 \leq i \leq L$$ → the main diagonal

$$S(i,i-1) = 0 \quad \forall \quad 2 \leq i \leq L$$ → the diagonal below

L: the length of input sequence

2. Recursion

Fill up the table (DP matrix) -- diagonal by diagonal

		G	G	G	A	A	A	U	C	C
	G	0	0	0	0					
	G	0	0	0	0	0				
	G		0	0	0	0	0			
	A			0	0	0	0	?		
	A				0	0	0	1		
	A					0	0	1	1	
	U						0	0	0	0
	C							0	0	0
	C								0	0

j →

i ↓

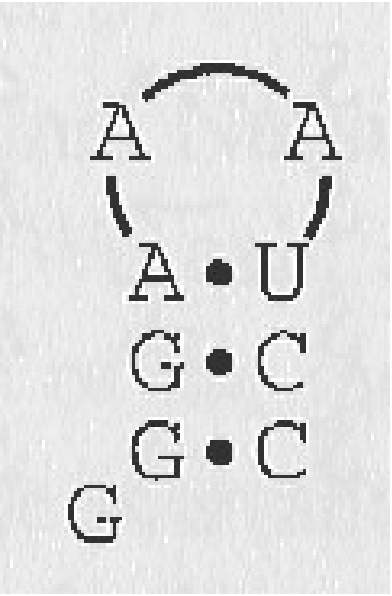
$$S(i,j) = max \begin{cases} S(i+1, j-1) + w(i,j) & (1) \\ S(i+1, j) & (2) \\ S(i, j-1) & (3) \\ max_{i<k<j} S(i,k) + S(k+1, j) & (4) \end{cases}$$

$$w(i,j) = \begin{cases} 1 & i, j \text{ are complementary} \\ 0 & otherwise \end{cases}$$

3. Traceback

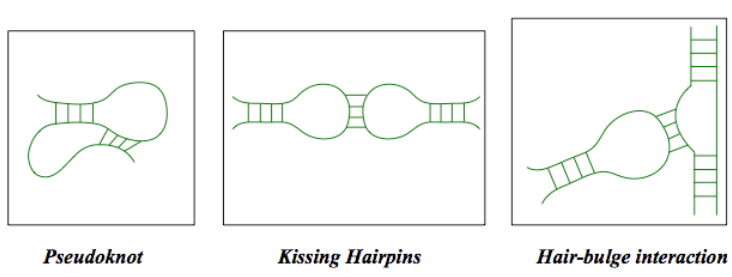
	G	G	G	A	A	A	U	C	C
G	0	0	0	0	0	0	1	2	3
G	0	0	0	0	0	0	1	2	3
G		0	0	0	0	0	1	2	2
A			0	0	0	0	1	1	1
A				0	0	0	1	1	1
A					0	0	1	1	1
U						0	0	0	0
C							0	0	0
C								0	0

The structure is:



What are the other “optimal” structures?

Pseudoknots drastically increase computational complexity



A simple extension to Nussinov's algorithm (DP) to include pseudoknots:

$$S_{i,j} = \max\{ \quad C_{i,j}, \quad \max_{i \le h_1 \le h_2 \le k < j_1 \le l_2 \le j} \{P_{i,k,h_1,h_2} + P_{k+1,j,l_1,l_2} + H_{h_1,h_2,l_1,l_2}\} \}$$

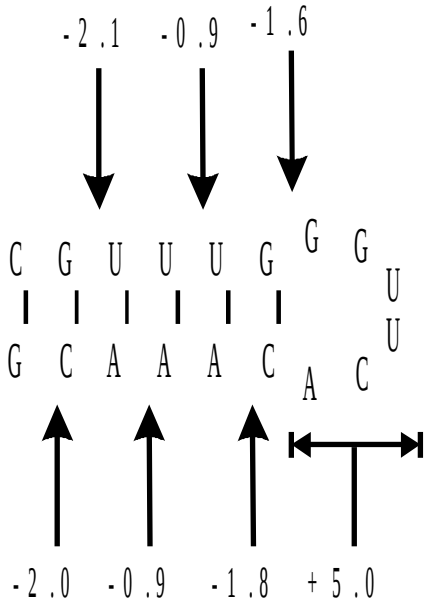
where P is a substructure containing a "dangling" base paring region:

$$P_{i,k,h_1,h_2} = \max\{ \max_{h_1 \le s < k} \{P_{i,s,h_1,h_2} + S_{s+1,k}\}, \max_{i \le t < h_1} \{S_{i,t} + P_{t+1,k,h_1,h_2}\} \}$$

Objective: Minimize Secondary Structure

Free Energy at 37 ºC:

Instead of δ(i, j), measure and sum energies:



$$\Delta G^\circ_{\text{helix}} = \Delta G^\circ \begin{matrix} C & G \\ G & C \end{matrix} + \Delta G^\circ \begin{matrix} G & U \\ C & A \end{matrix} + 2\Delta G^\circ \begin{matrix} U & U \\ A & A \end{matrix} + \Delta G^\circ \begin{matrix} U & G \\ A & C \end{matrix} =$$

-2.0 kcal/mol - 2.1 kcal/mol + 2x(-0.9) kcal/mol - 1.8 kcal/mol = -7.7 kcal/mol

$$\Delta G^\circ_{\text{hairpin loop}} = \Delta G^\circ_{\text{initiation}} \text{ (6 nucleotides)} + \Delta G^\circ_{\text{mismatch}} \begin{matrix} G & G \\ C & A \end{matrix} =$$

5.0 kcal/mol - 1.6 kcal/mol = 3.4 kcal/mol

$$\Delta G^\circ_{\text{total}} = \Delta G^\circ_{\text{hairpin}} + \Delta G^\circ_{\text{helix}} = 3.4 \text{ kcal/mol} - 7.7 \text{ kcal/mol} = -4.3 \text{ kcal/mol}$$

Mathews, Disney, Childs, Schroeder, Zuker, & Turner. 2004. *PNAS 101:* 7287.

Zuker's algorithm MFOLD: computing loop dependent energies

$$W_{i,j} = \min\{W_{i+1,j},\ W_{i,j-1},\ V_{i,j},\ \min_{i \le k < j}\{W_{i,k} + W_{k+1,j}\}\}$$

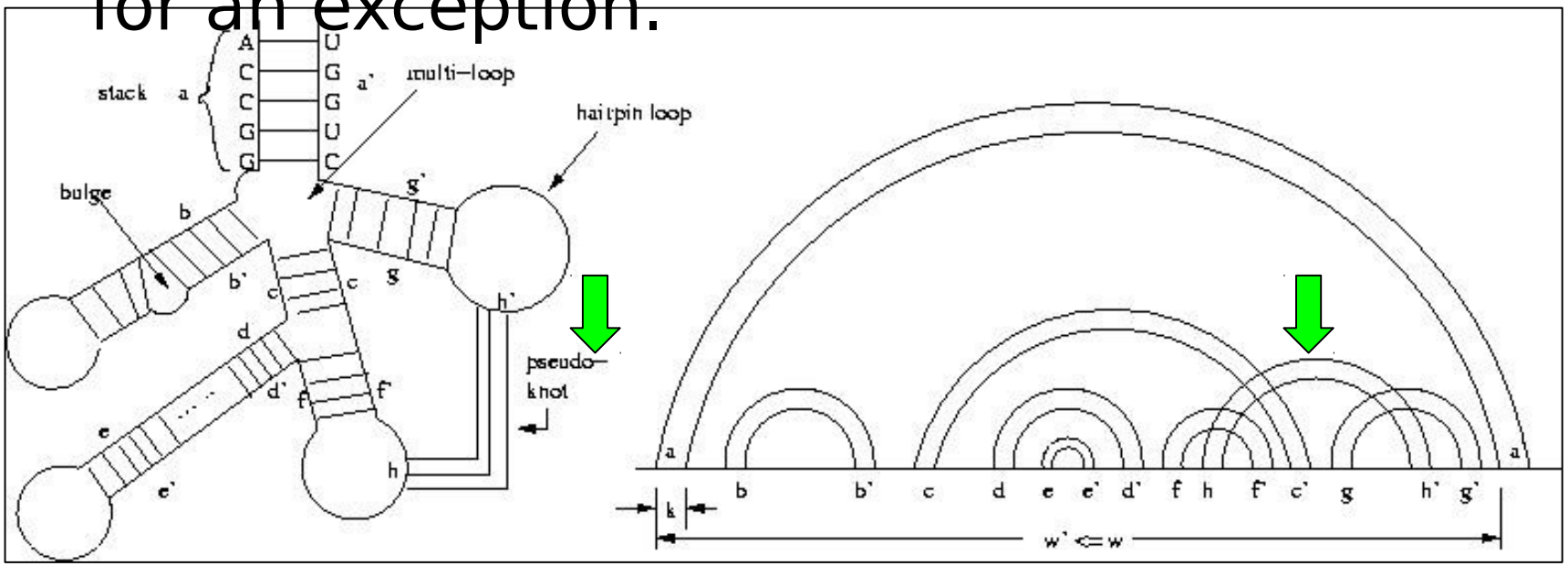
$$V_{i,j} = \min\{H_{i,j},\ S_{i,j},\ VBI_{i,j},\ VM_{i,j}\}$$

$$VBI_{i,j} = \min_{i < i' < j < j'}\{I_{i,j,i',j'} + V_{i',j'}\}$$

$$VM_{i,j} = \min_{i+1 < k < j-1}\{W_{i+1,k} + W_{k+1,j-1}\}$$

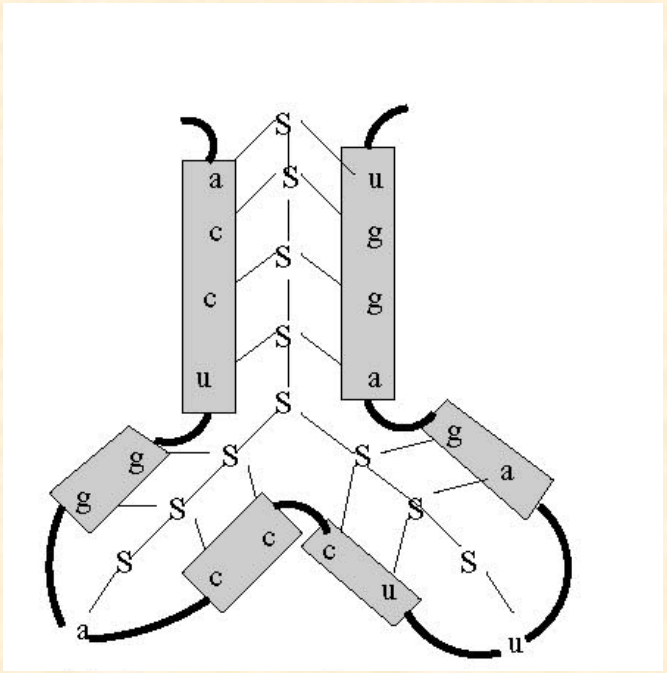
RNA structure

- Base-pairing defines a secondary structure. The base-pairing is usually non-crossing.
- In this example the pseudo-knot makes for an exception.



Stochastic context-free grammar

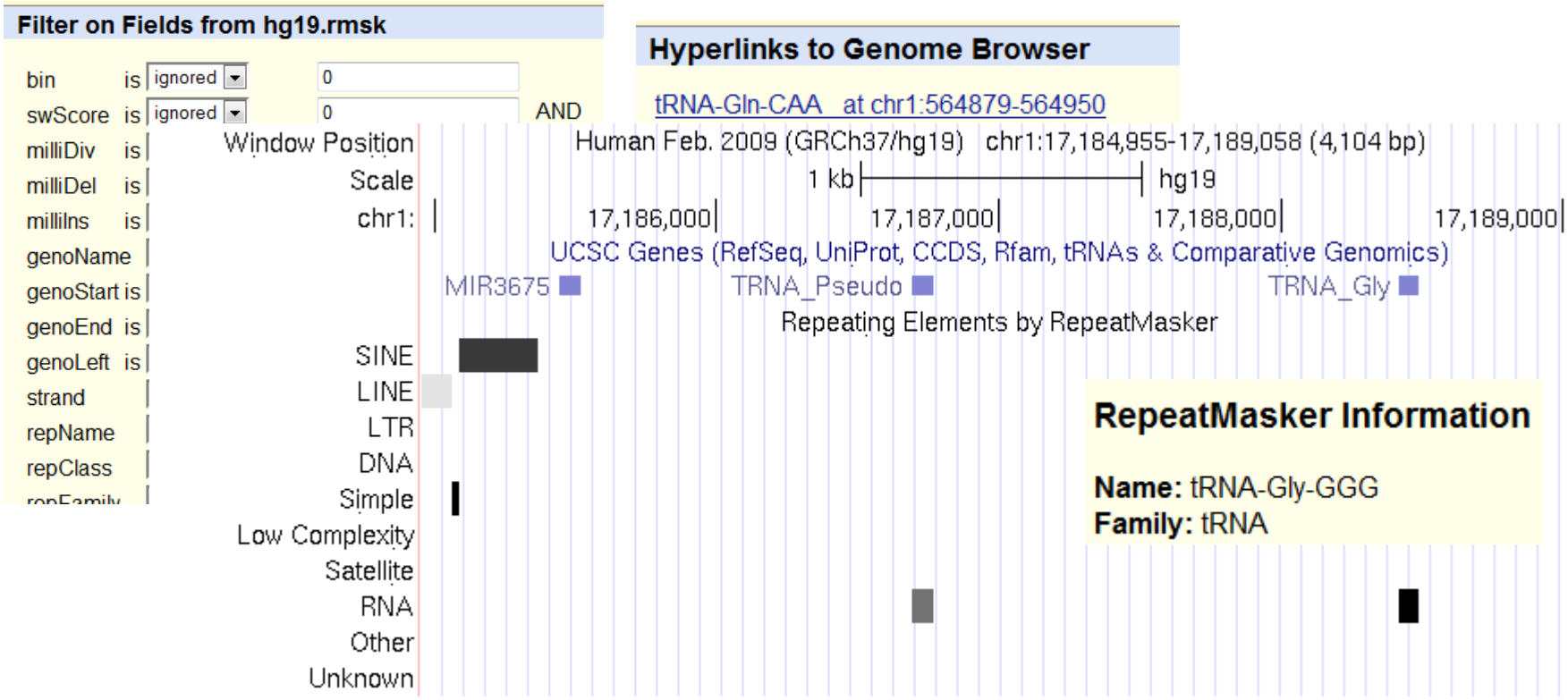
1. A CFG	2. A derivation of “accuggacccuuagaggu”
S → aSu	S → aSu
S → cSg	→ acSgu
S → gSc	→ accSggu
S → uSa	→ accuSaggu
S → a	→ accuSSaggu
S → c	→ accugScSaggu
S → g	→ accuggSccSaggu
S → u	→ accuggaccSaggu
S → SS	→ accuggacccSgaggu
	→ accuggacccuSagaggu
	→ accuggacccuuagaggu



3. Corresponding structure

ssRNA transcription

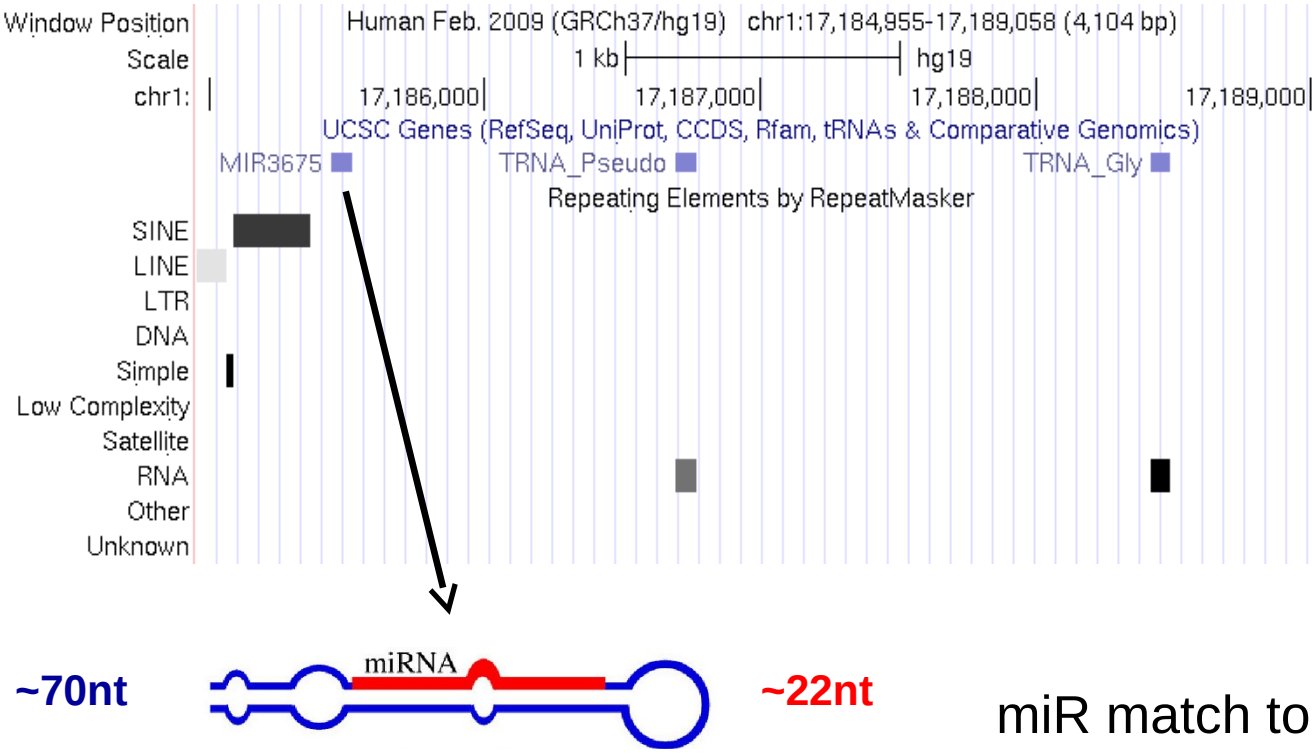
- ssRNAs like tRNAs are usually encoded by short "non coding" genes, that transcribe independently.
- Found in both the UCSC "known genes" track, and as a subtrack of the RepeatMasker track



“non coding” RNAs (ncRNA) subtype:

microRNAs (miRNA/miR)

MicroRNA (miR)

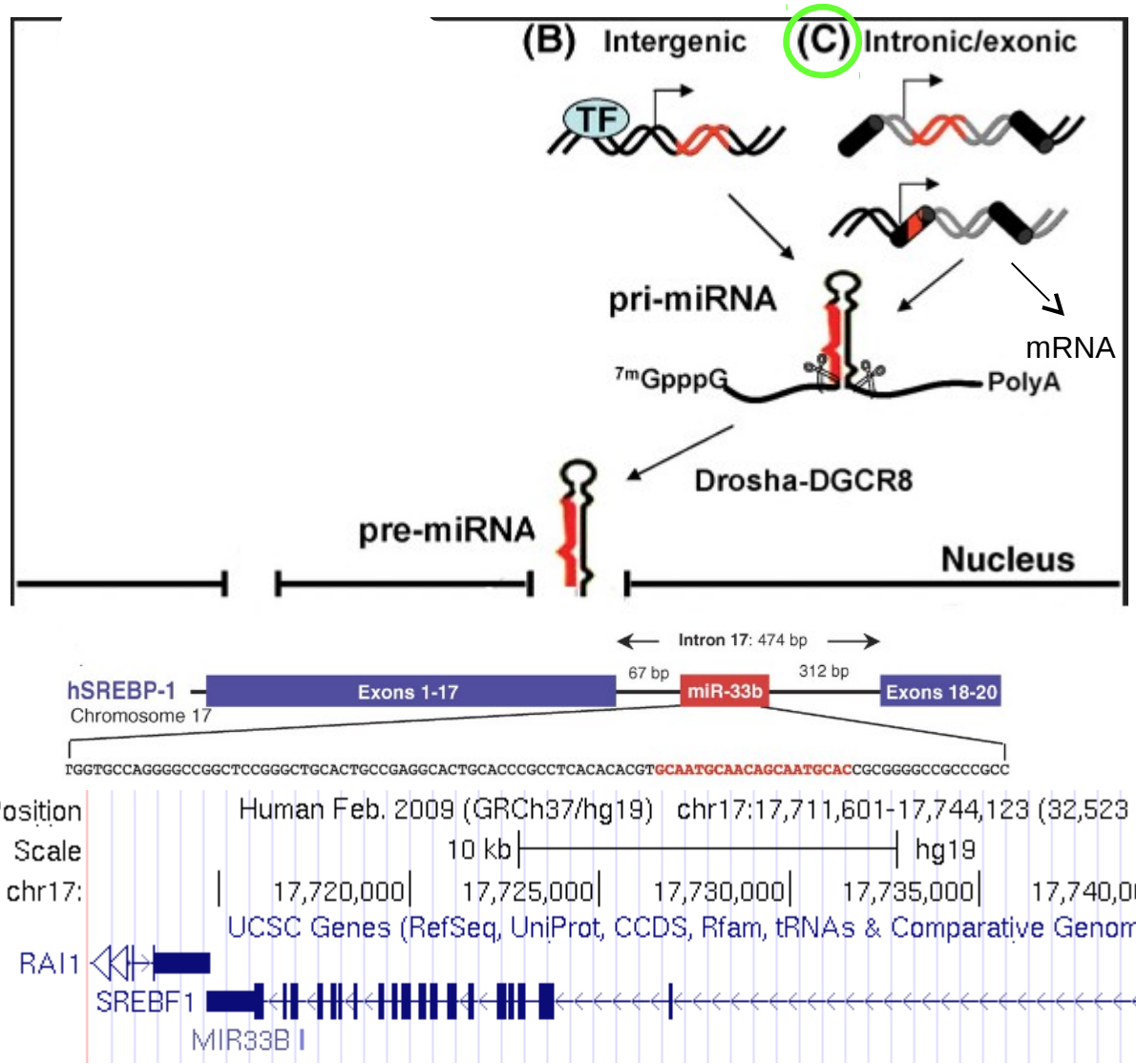


miR match to target mRNA is quite loose.

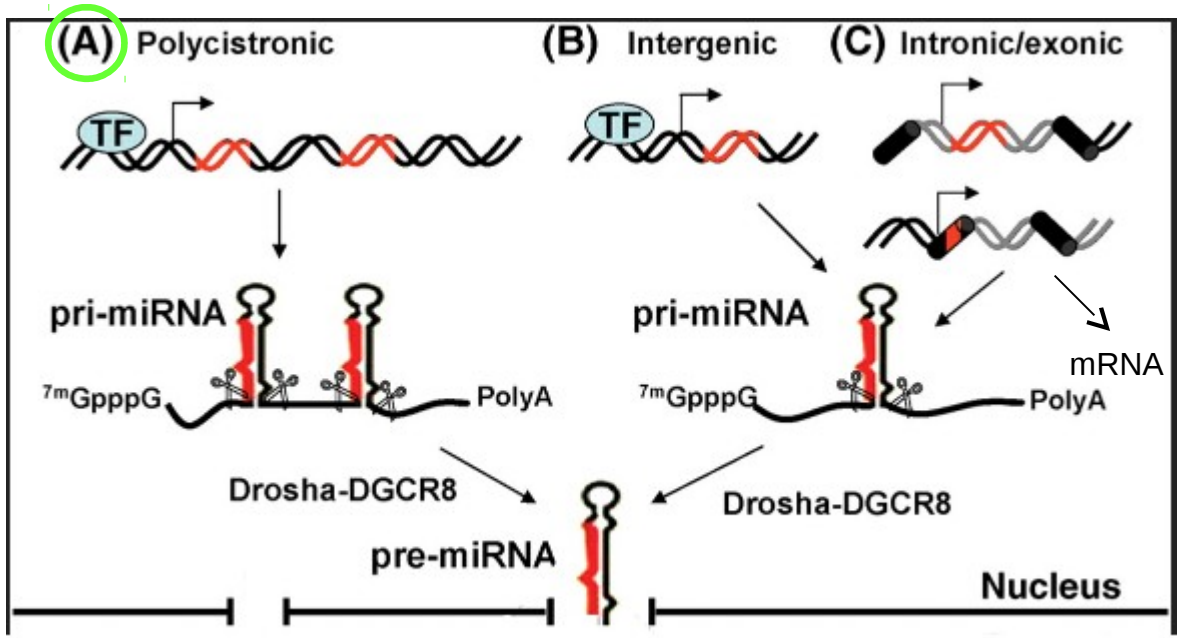
```
3' uagcgccaaauauggUUUACUUA 5' has-miR-579
   ||: |  |||||:  ||||||:|
5' atttctttttatggaAAATGAGT 3' LR1G3
     Out-seed        Seed
```

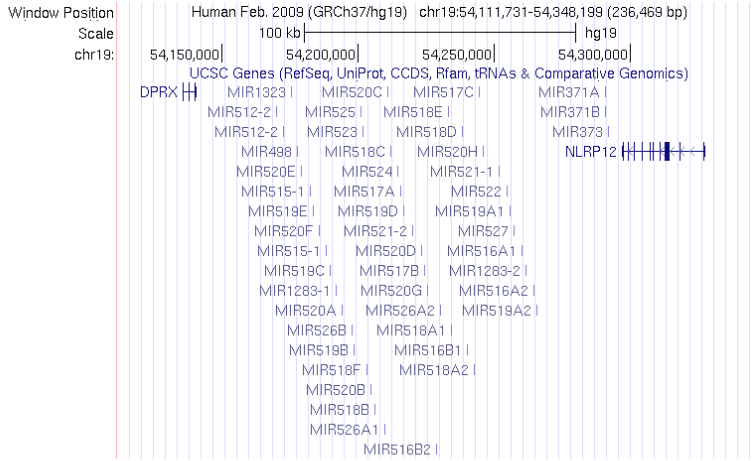
 a single miR can regulate the expression of hundreds of genes.

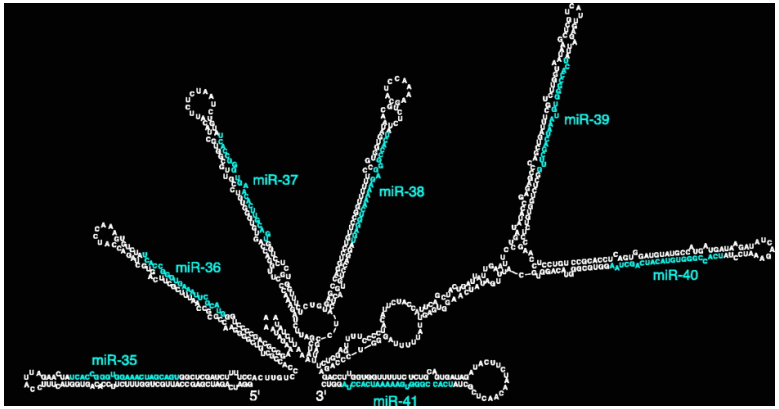
MicroRNA Transcription



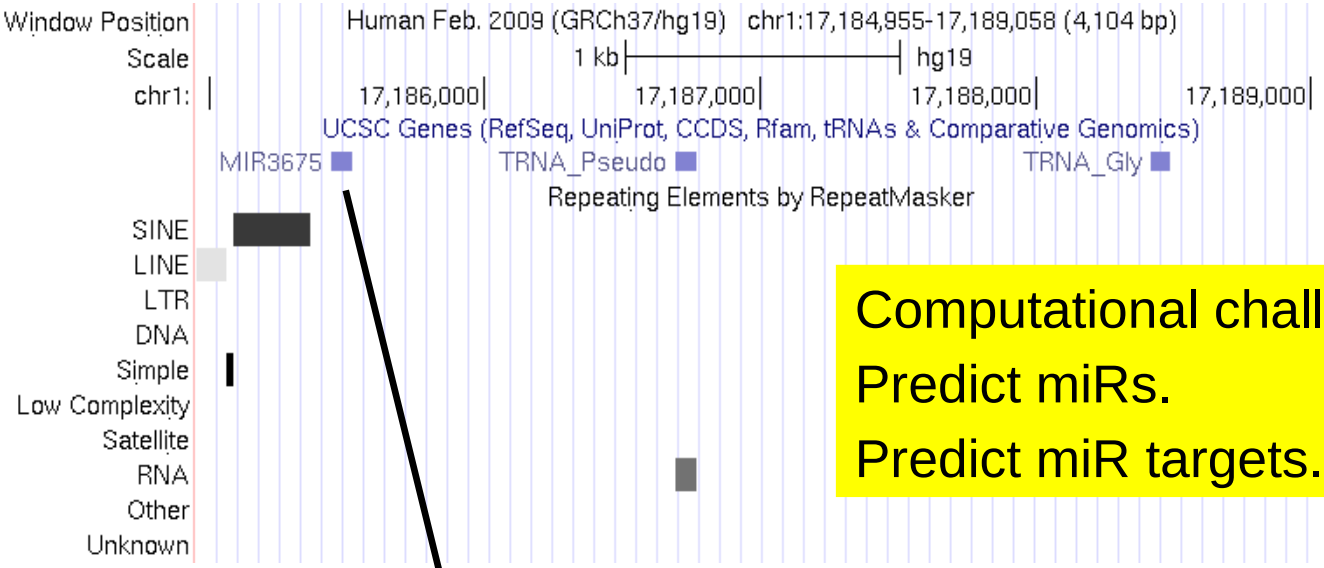
MicroRNA Transcription







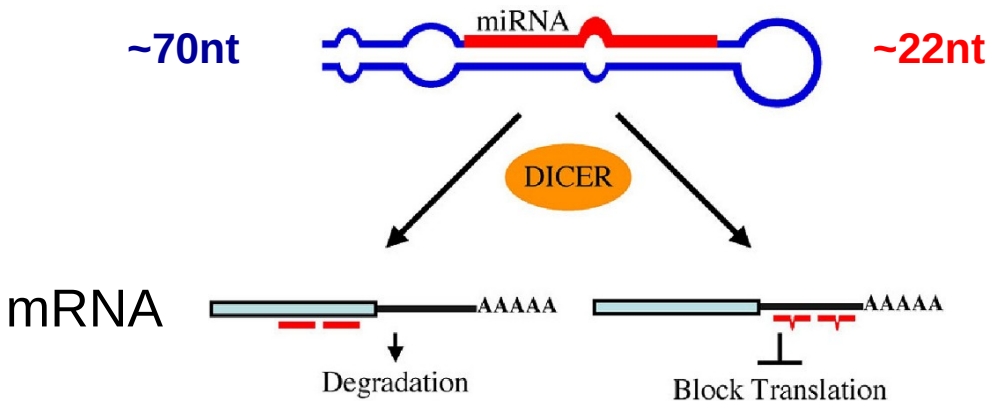
MicroRNA (miR)



Computational challenges:

Predict miRs.

Predict miR targets.



miR match to target mRNA is quite loose.

```
3' uagcgccaaauauggUUUACUUA 5' has-miR-579
   ||: |  |||||:  |||||||:|
5' atttctttttatggaAAATGAGT 3' LR1G3
     Out-seed        Seed
```

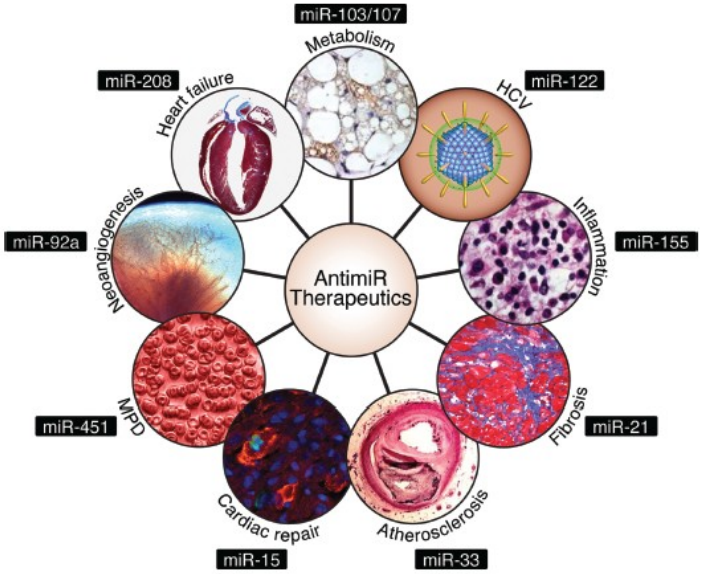
- a single miR can regulate the expression of hundreds of genes.

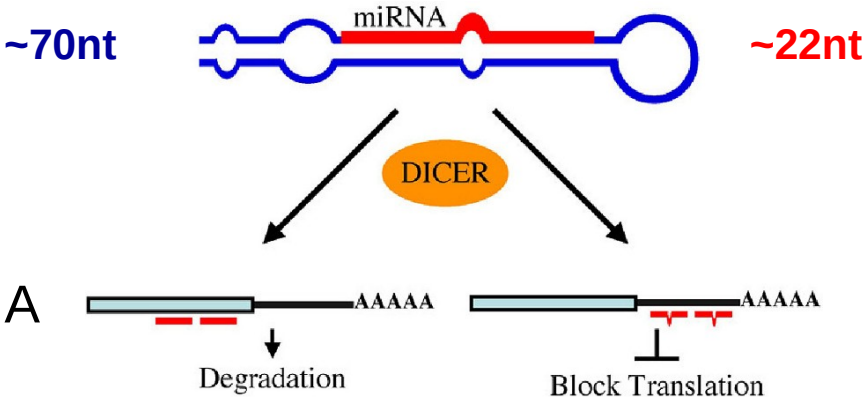
MicroRNA Therapeutics

Idea: bolster/inhibit miR production to broadly modulate protein production

Hope: “right” the good guys and/or “wrong” the bad guys

Challenge: and not vice versa.





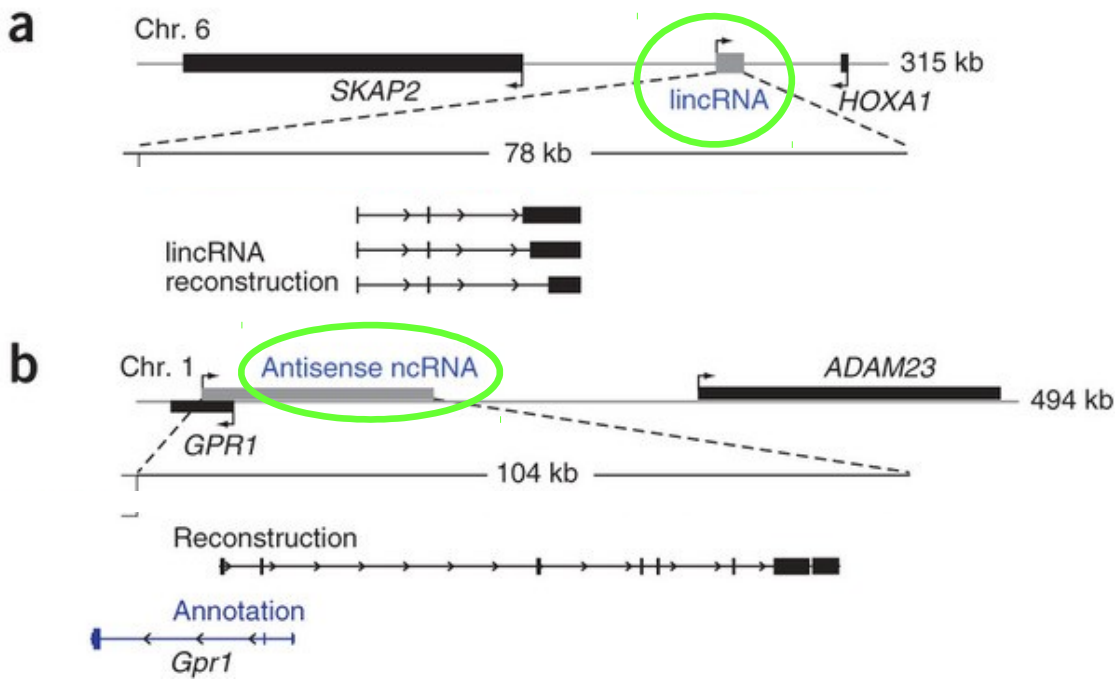
miR match to target mRNA is quite loose.

```
3' uagcgccaaauauggUUUACUUA 5' has-miR-579
   ||: |  |||||:  |||||||:|
5' atttctttttatggaAAATGAGT 3' LR1G3
    Out-seed        Seed
```

⇒ a single miR can regulate the expression of hundreds of genes.

Other Non Coding Transcripts

lncRNAs (long non coding RNAs)



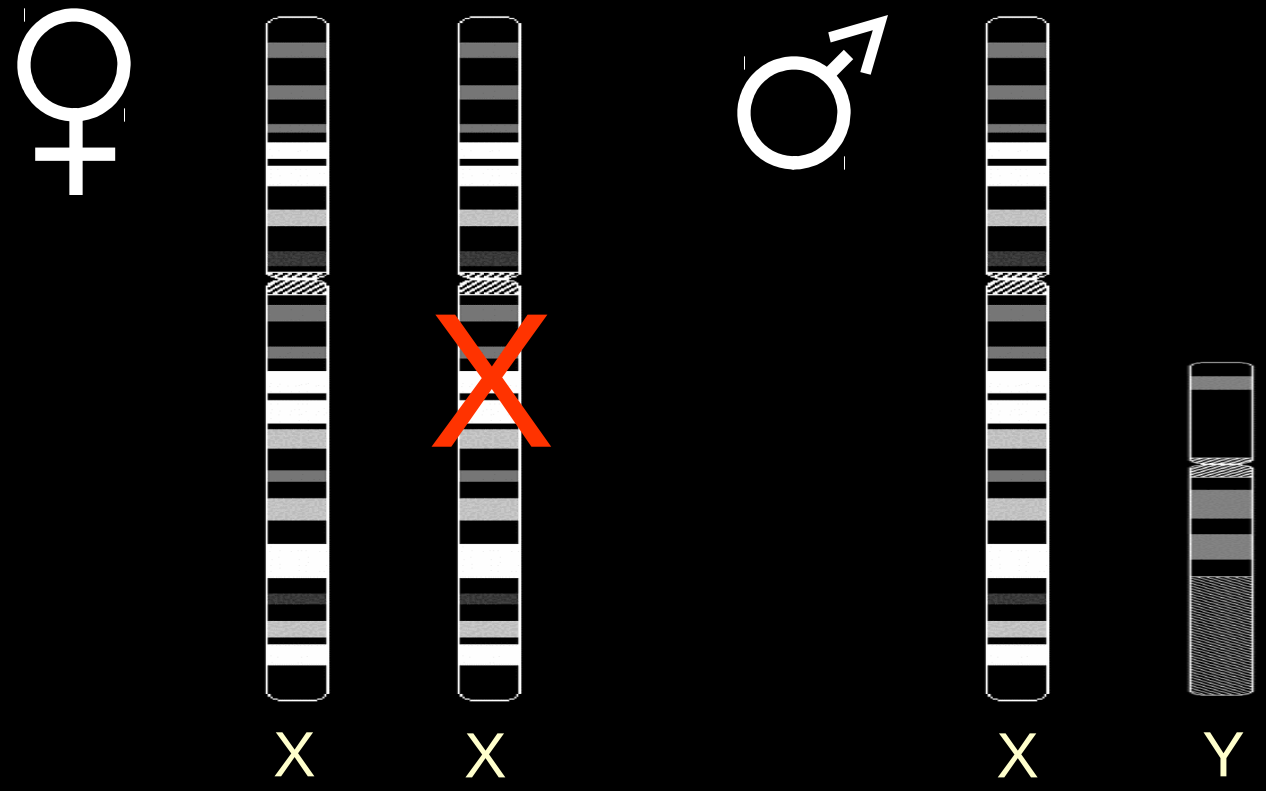
Don’t seem to fold into clear structures (or only a sub-region does).

Diverse roles only now starting to be understood.

Example: bind splice site, cause intron retention.

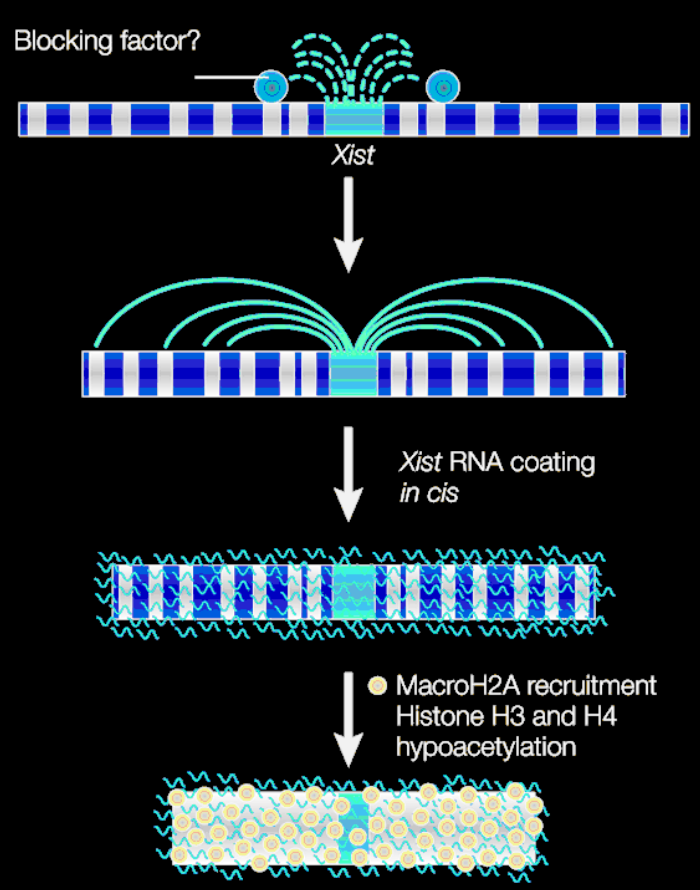
 Challenges: 1) detect and 2) predict function computationally

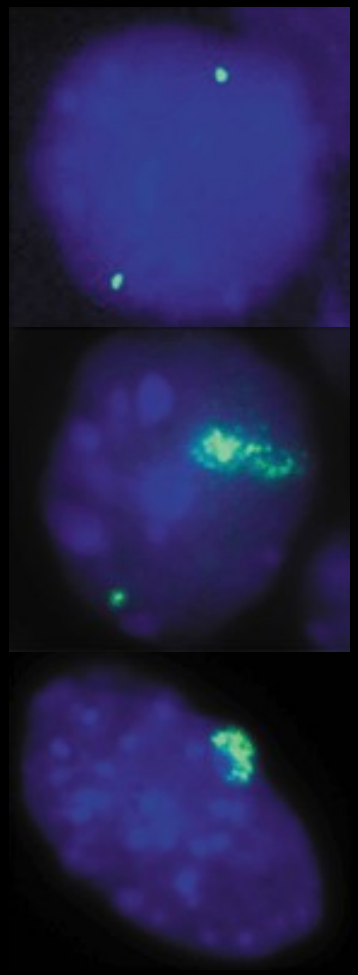
X chromosome inactivation in mammals



Dosage compensation

Xist – X inactive-specific transcript





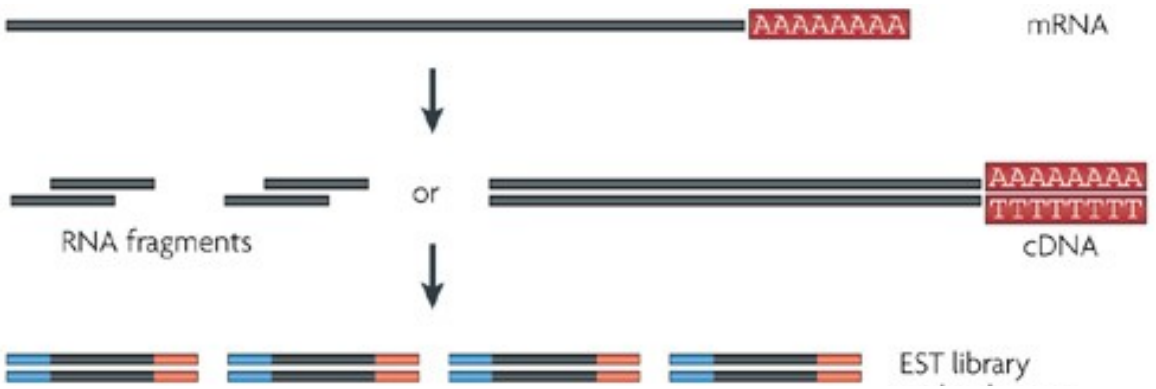
Avner and Heard, Nat. Rev. Genetics 2001 2(1):59-67

System output measurements

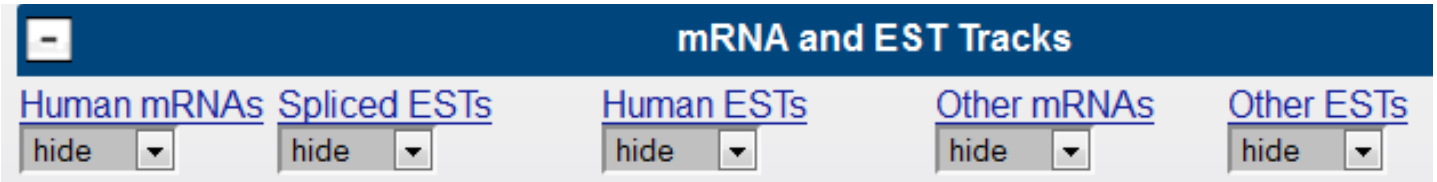
We can measure non/coding gene expression!

(just remember – it is context dependent)

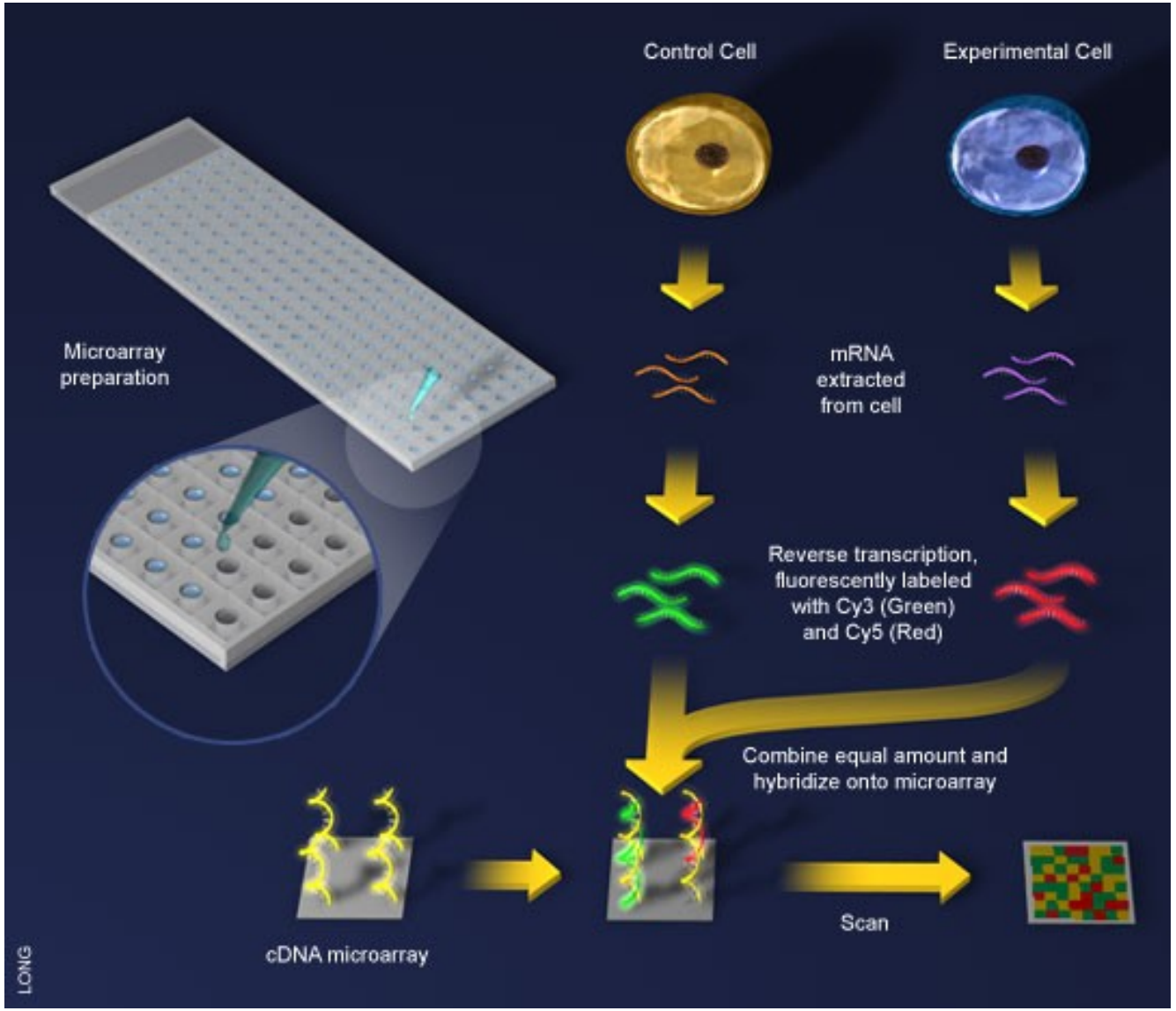
1. First generation mRNA (cDNA) and EST sequencing:



In UCSC Browser:

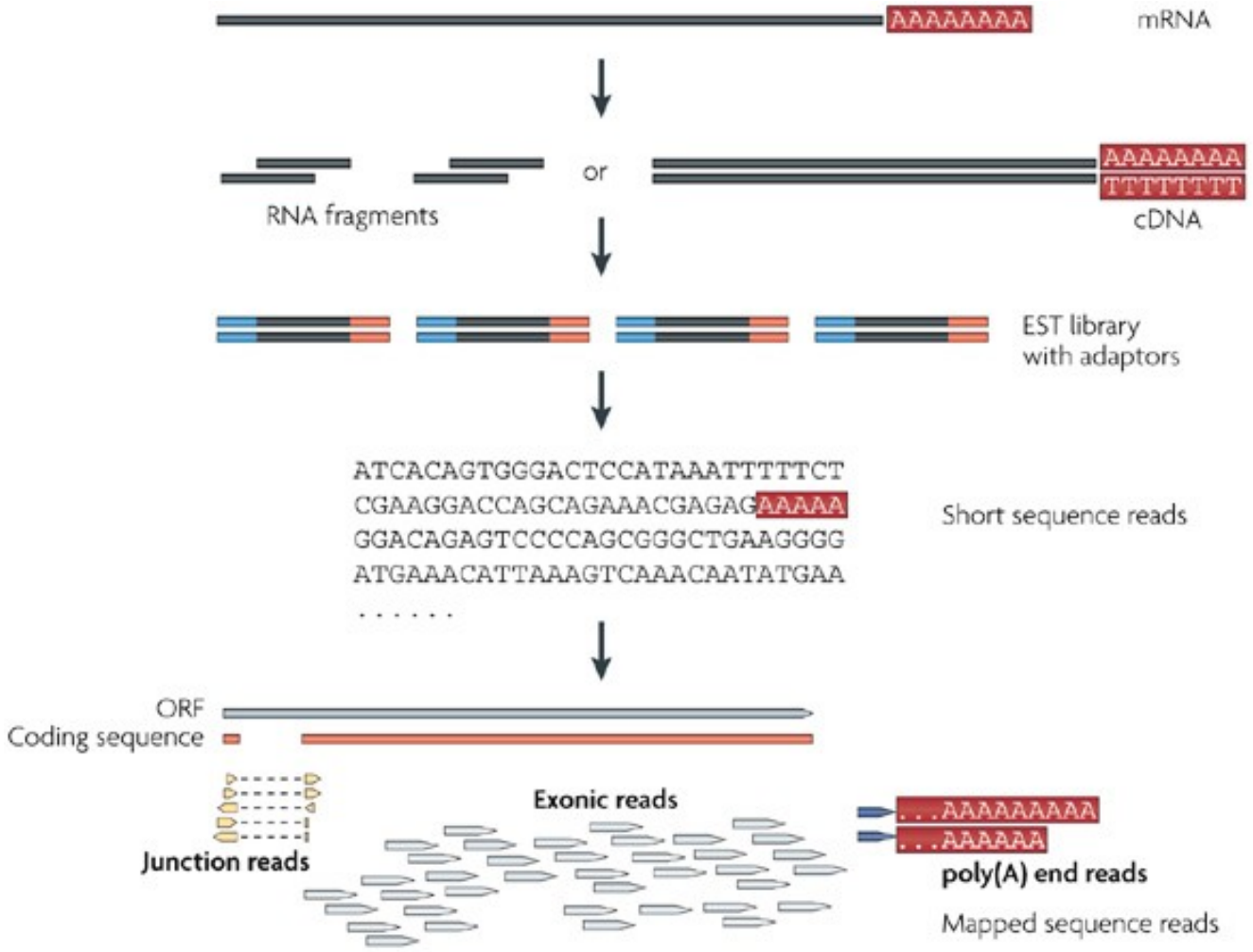


2. Gene Expression Microarrays (“chips”)

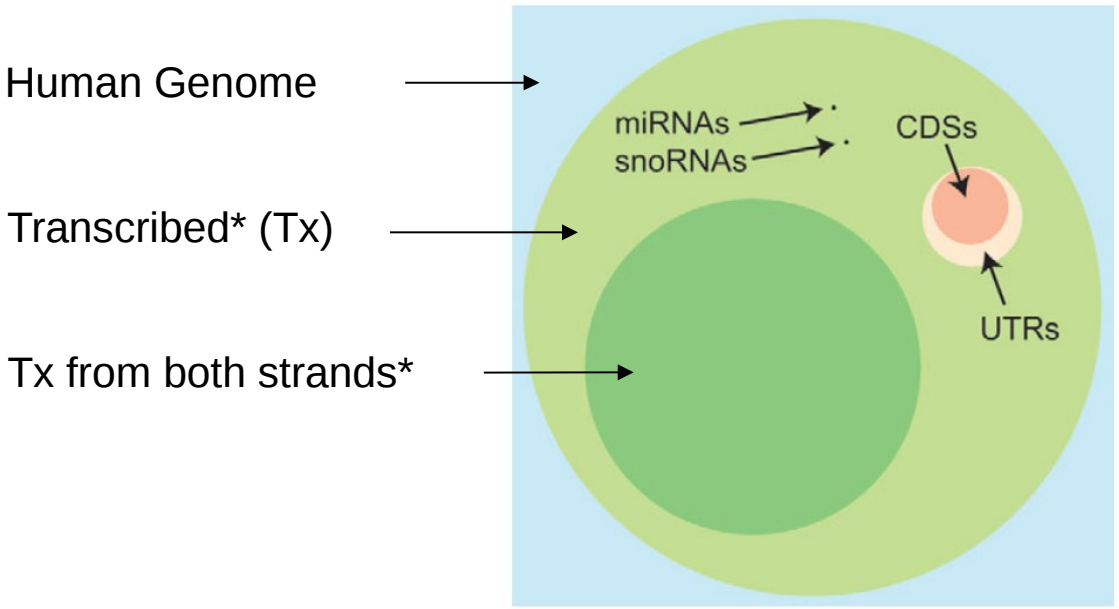


3. RNA-seq

"Next" (2nd) generation sequencing.



Transcripts, transcripts *everywhere*



* True size of set unknown

Or are they?

Most "Dark Matter" Transcripts Are Associated With Known Genes

Harm van Bakel[1], Corey Nislow[1,2], Benjamin J. Blencowe[1,2], Timothy R. Hughes[1,2]*

1 Banting and Best Department of Medical Research, University of Toronto, Toronto, Ontario, Canada, **2** Department of Molecular Genetics, University of Toronto, Toronto, Ontario, Canada

Abstract

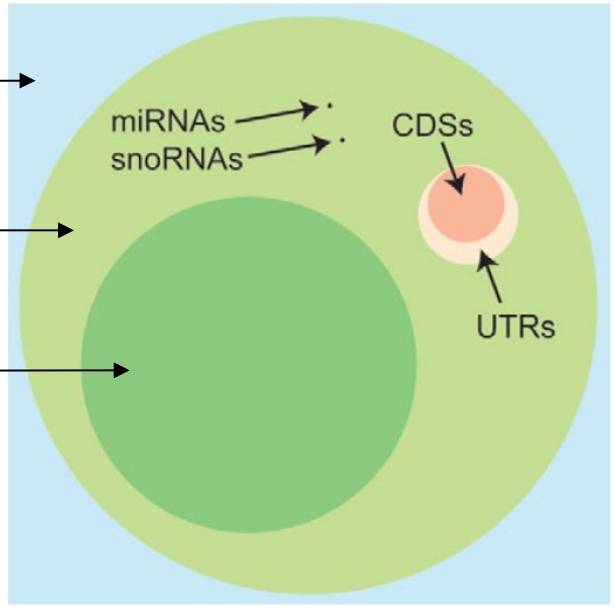
A series of reports over the last few years have indicated that a much larger portion of the mammalian genome is transcribed than can be accounted for by currently annotated genes, but the quantity and nature of these additional transcripts remains unclear. Here, we have used data from single- and paired-end RNA-Seq and tiling arrays to assess the quantity and composition of transcripts in PolyA+ RNA from human and mouse tissues. Relative to tiling arrays, RNA-Seq identifies many fewer transcribed regions ("seqfrags") outside known exons and ncRNAs. Most nonexonic seqfrags are in introns, raising the possibility that they are fragments of pre-mRNAs. The chromosomal locations of the majority of intergenic seqfrags in RNA-Seq data are near known genes, consistent with alternative cleavage and polyadenylation site usage, promoter- and terminator-associated transcripts, or new alternative exons; indeed, reads that bridge splice sites identified 4,544 new exons, affecting 3,554 genes. Most of the remaining seqfrags correspond to either single reads that display characteristics of random sampling from a low-level background or several thousand small transcripts (median length = 111 bp) present at higher levels, which also tend to display sequence conservation and originate from regions with open chromatin. We conclude that, while there are bona fide new intergenic transcripts, their number and abundance is generally low in comparison to known exons, and the genome is not as pervasively transcribed as previously reported.

The million dollar question

Human Genome

Transcribed* (Tx)

Tx from both strands*



Leaky tx?

Functional?

* True size of set unknown

Gene Finding II: *technology dependence*

Challenge:

"Find the genes, the whole genes, and nothing but the genes"

We started out trying to predict genes directly from the genome.

When you measure gene expression, the challenge changes:
Now you want to build gene models from your observations.

These are both technology dependent challenges.

The hybrid: what we measure is a tiny fraction of the space-time state space for cells in our body. We want to generalize from measured states and improve our predictions for the full compendium of states.