CSE 428
CompBio Capstone
Spring 2012

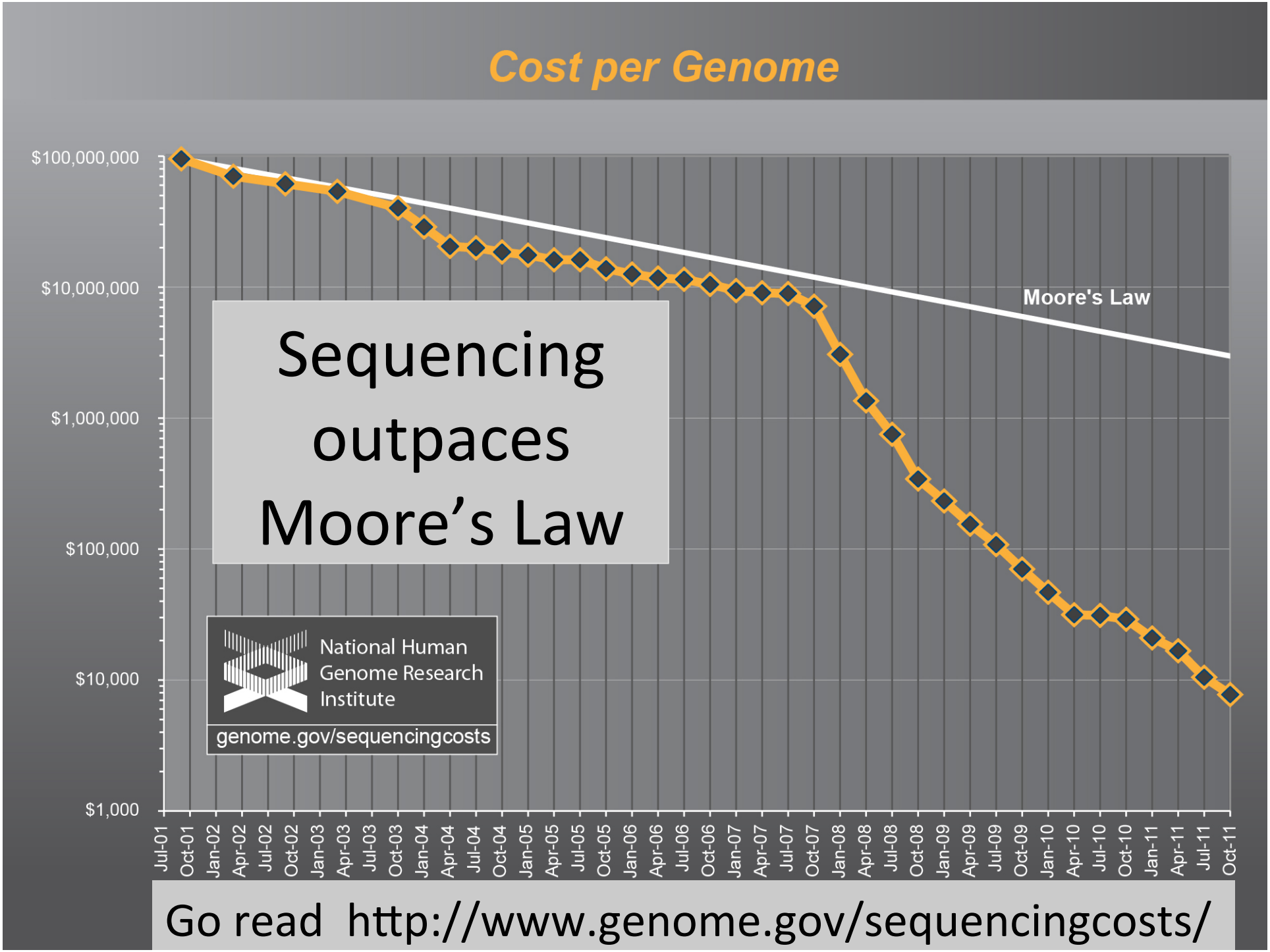
Intro

Larry Ruzzo

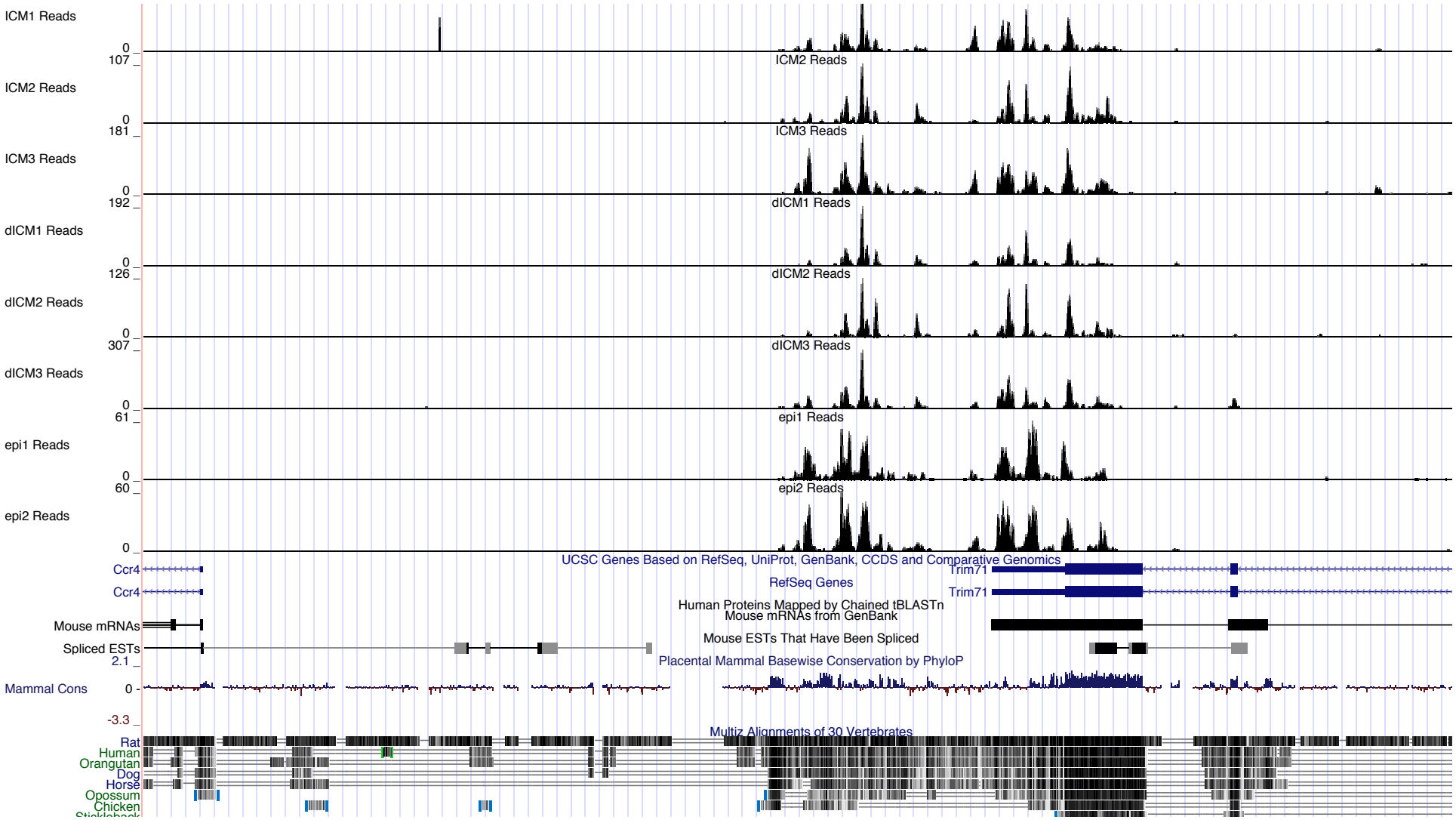
2 Revolutionary Changes in Bio

1. Quantum leaps in sequencing technology

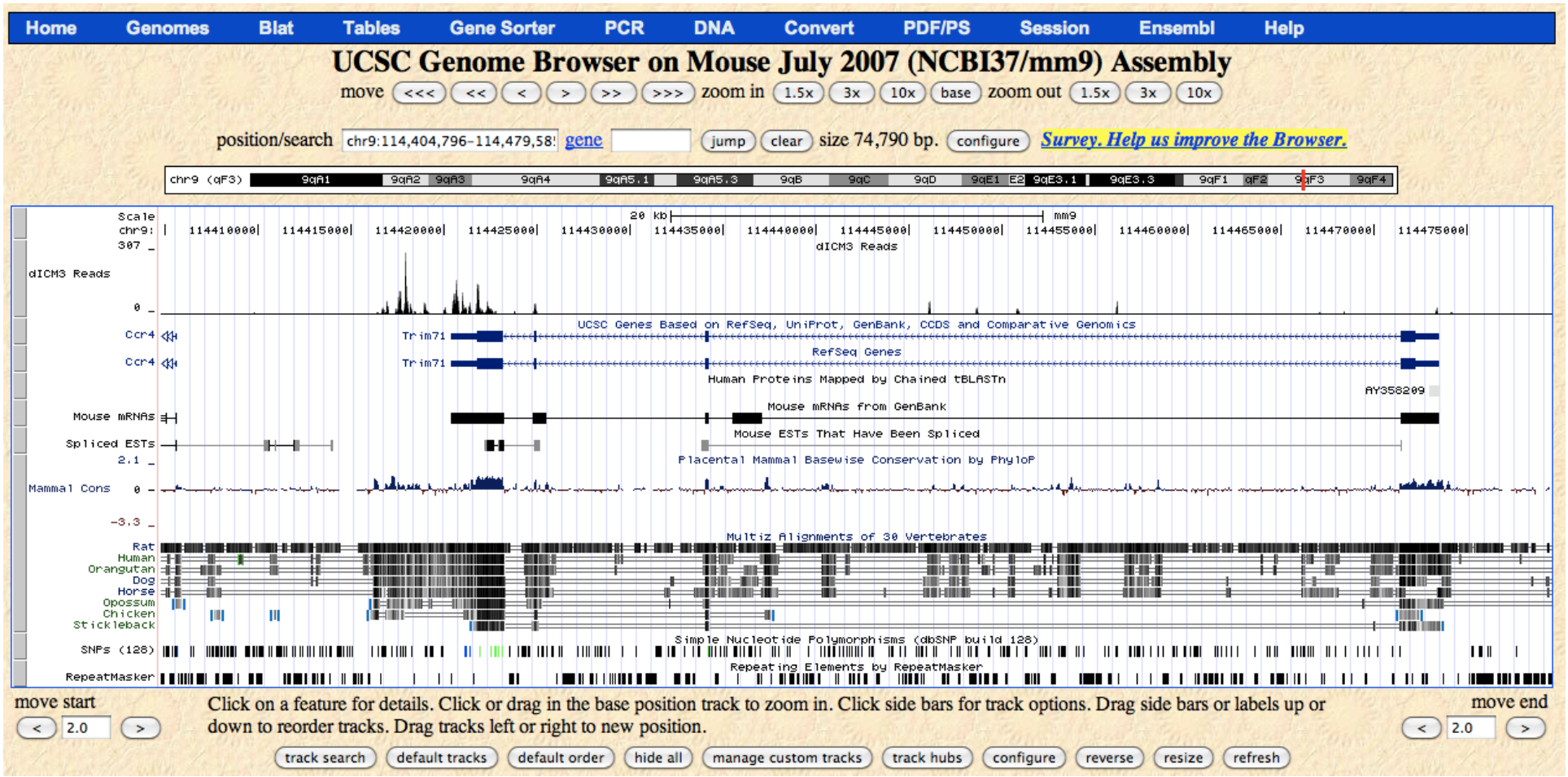
2. Widespread non-protein-coding RNA


Cost per Genome
$100,000,000
$10,000,000
$1,000,000
$100,000
$10,000
$1,000
Moore's Law
Sequencing outpaces Moore's Law
National Human Genome Research Institute
genome.gov/sequencingcosts
Jul-01
Oct-01
Jan-02
Apr-02
Jul-02
Oct-02
Jan-03
Apr-03
Jul-03
Oct-03
Jan-04
Apr-04
Jul-04
Oct-04
Jan-05
Apr-05
Jul-05
Oct-05
Jan-06
Apr-06
Jul-06
Oct-06
Jan-07
Apr-07
Jul-07
Oct-07
Jan-08
Apr-08
Jul-08
Oct-08
Jan-09
Apr-09
Jul-09
Oct-09
Jan-10
Apr-10
Jul-10
Oct-10
Jan-11
Apr-11
Jul-11
Oct-11
Go read http://www.genome.gov/sequencingcosts/

Some RNAseq Data



wider view of same locus



RNA Discovery

lots more cool stuff

Dogma: DNA $\xrightarrow{>>2\%}$ RNA $\xrightarrow{<2\%}$ Protein → ~~all~~ the cool stuff

Evolutionarily *shared* structure ≈ non-noise

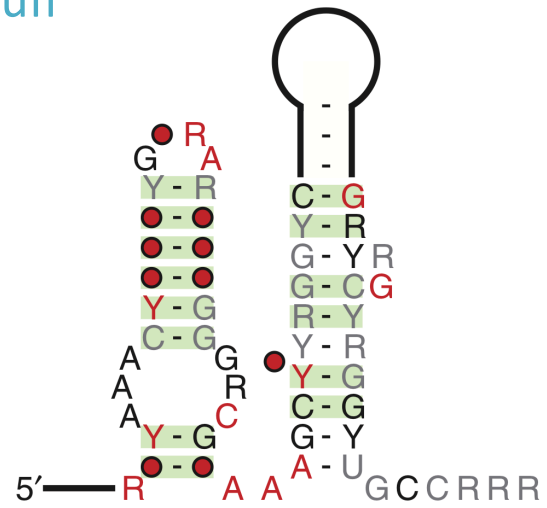
RNA structure: $O(n^3)$

big genomes, shared structure: even slower…

New Tools (Weinberg, Yao, Tseng)

better/faster search & discovery

⇒Discoveries:



Sudarsan, *et al., Science,* 2008

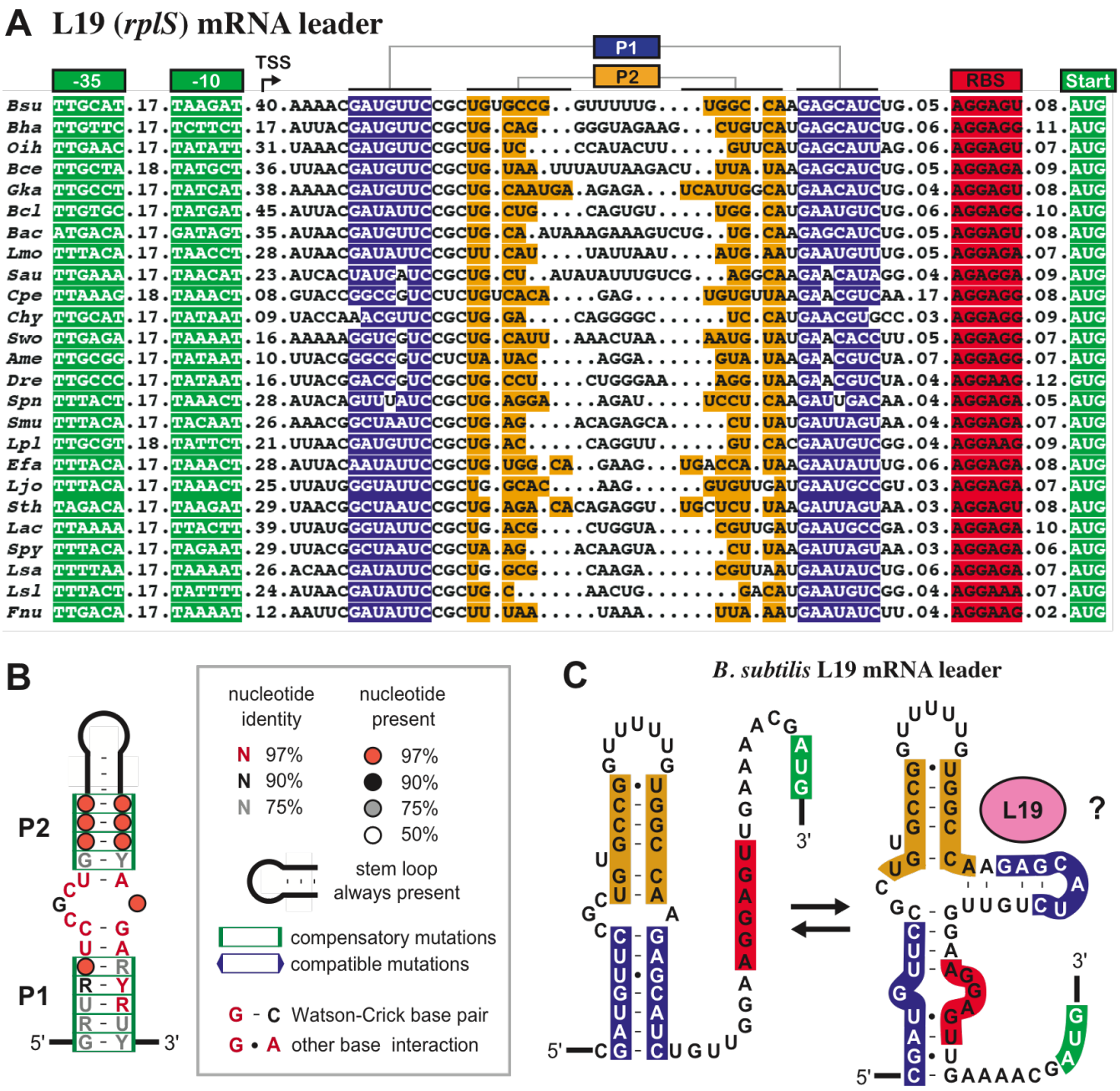
bacteria: ~ ⅓-½ of all known “riboswitches” (a few CPU-yrs)

humans: ~ 1000’s of candidates, a few verified (~250 CPU-yrs)

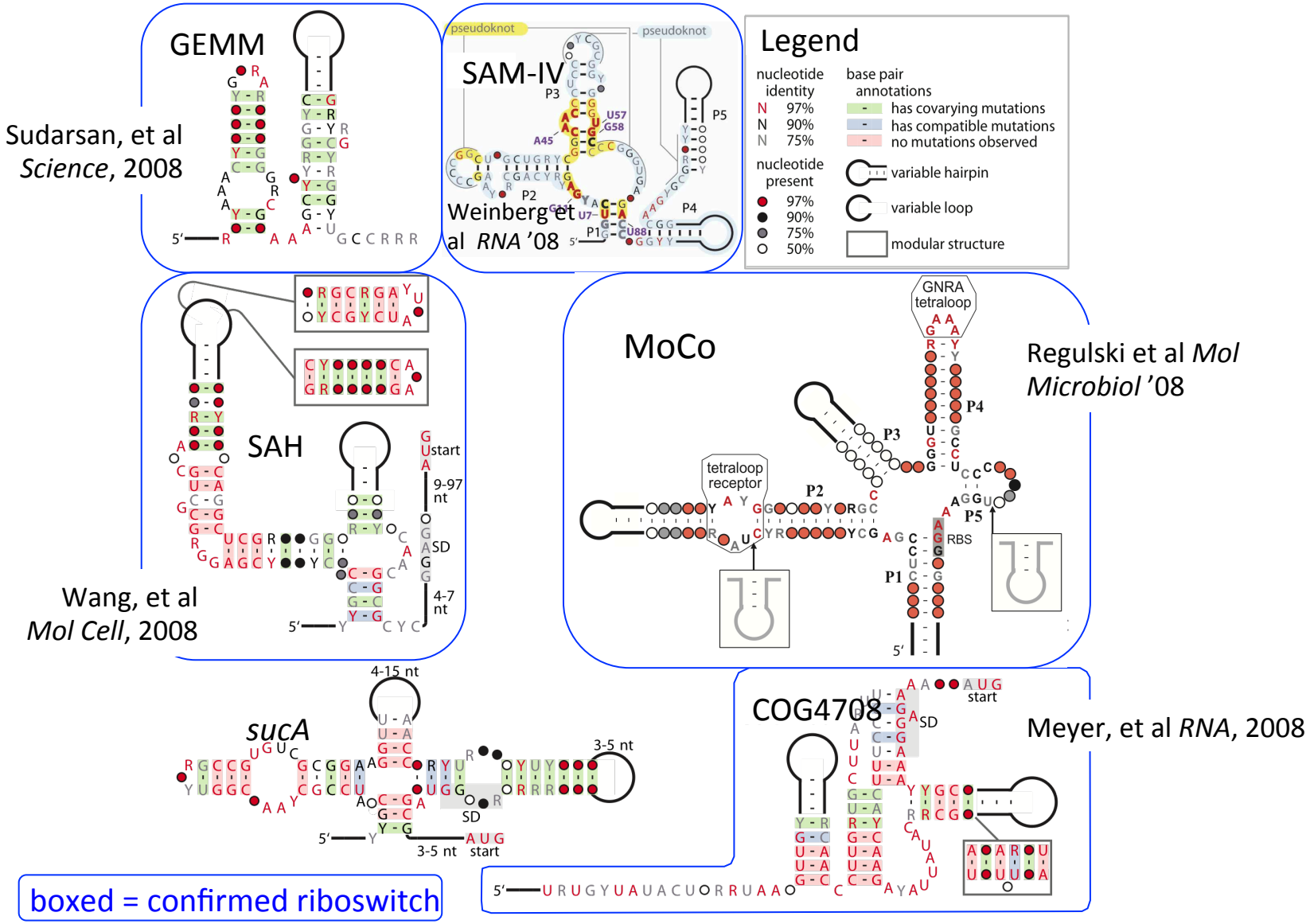
Future: speed, accuracy, data integration, biological validation

Example: Ribosomal Auto-regulation

Excess L19 represses L19 (RF00556; 555-559 similar)



Representative motifs



Weinberg, Barrick, Yao, Roth, Kim, Gore, Wang, Lee, Block, Sudarsan, Neph, Tompa, Ruzzo and Breaker. Identification of 22 candidate structured RNAs in bacteria using the CMfinder comparative genomics pipeline. *Nucl. Acids Res.*, July 2007 35: 4809-4819.

A Problem

Widespread RNA transcription is dismissed by many as “transcriptional noise” – rapidly degraded, functionally meaningless, etc.

Hypothesis

Functional noncoding RNA will display a greater level of cross-species conservation than expected by chance, where "conservation" includes preservation of RNA secondary structure, not just primary sequence.

The goal of the 428 project is to test this hypothesis.

Project Outline

- Phase 1:
  - IN: RNA seq raw data
  - map to genome
    - including spliced
  - identify (differentially) transcribed regions, splicing
  - crossref to published gene annotations
  - OUT: transcribed (spliced) noncoding loci

Project Outline (cont.)

- Phase 2:
  - IN: transcribed (spliced) noncoding loci
  - find aligned regions in other species
    - after splicing
  - multi-species RNA secondary structure prediction
    - and scoring
  - statistics: repeat above on “control” data,
    - “similar” non-transcribed regions, and/or
    - randomly shuffled versions of the “real” data
  - OUT: is there a statistically significant difference

tools

- ~~svn~~/git
- python/biopython
- R
- vienna RNA
- cmfinder
- infernal
- multiperm
- bwa/gsnap/bowtie/tophat/cufflinks?
- sql?
- liftover