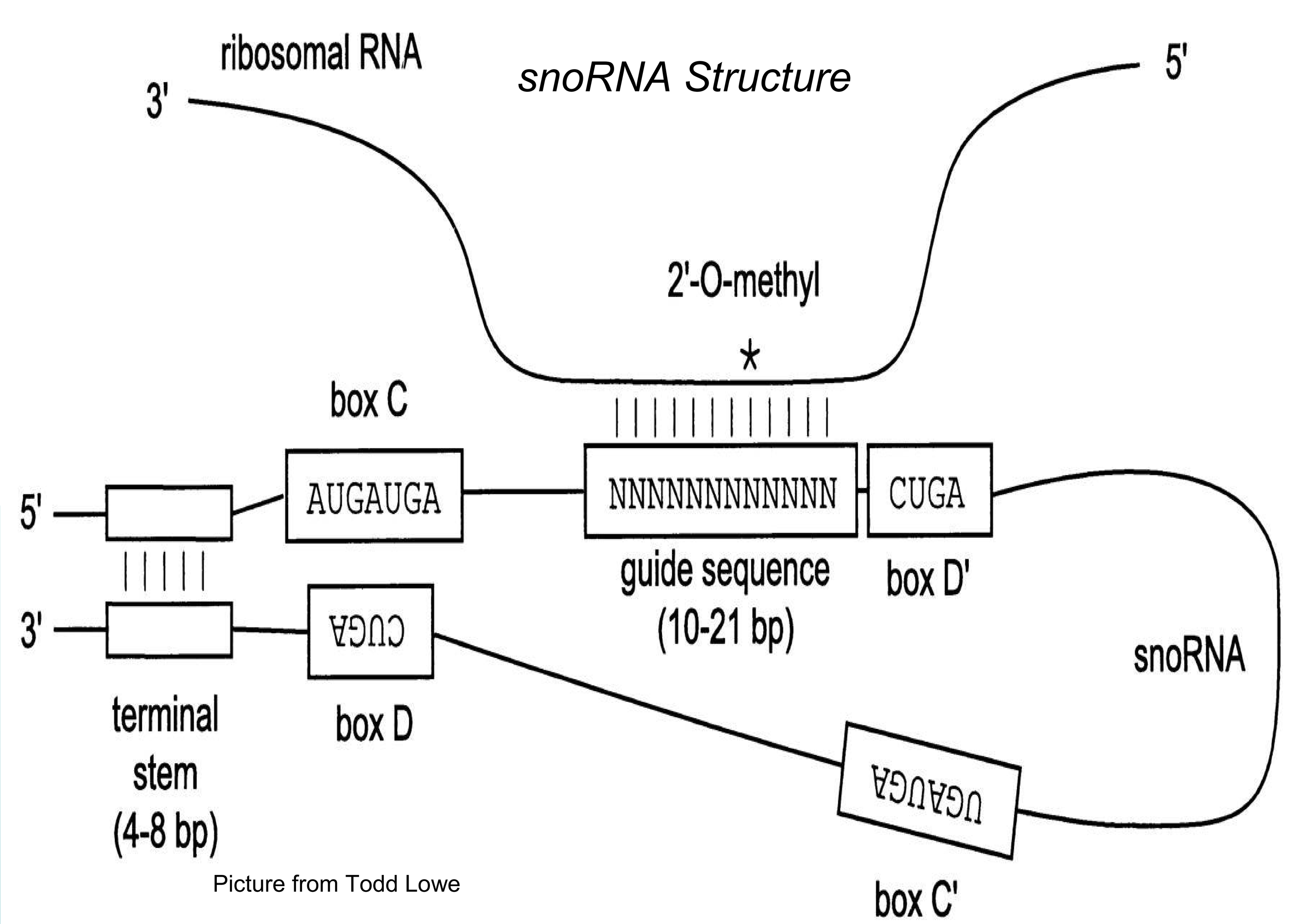
Using Conserved Elements to Assist in Gene Annotation

Christoph Rau

Todd Lowe, Advisor

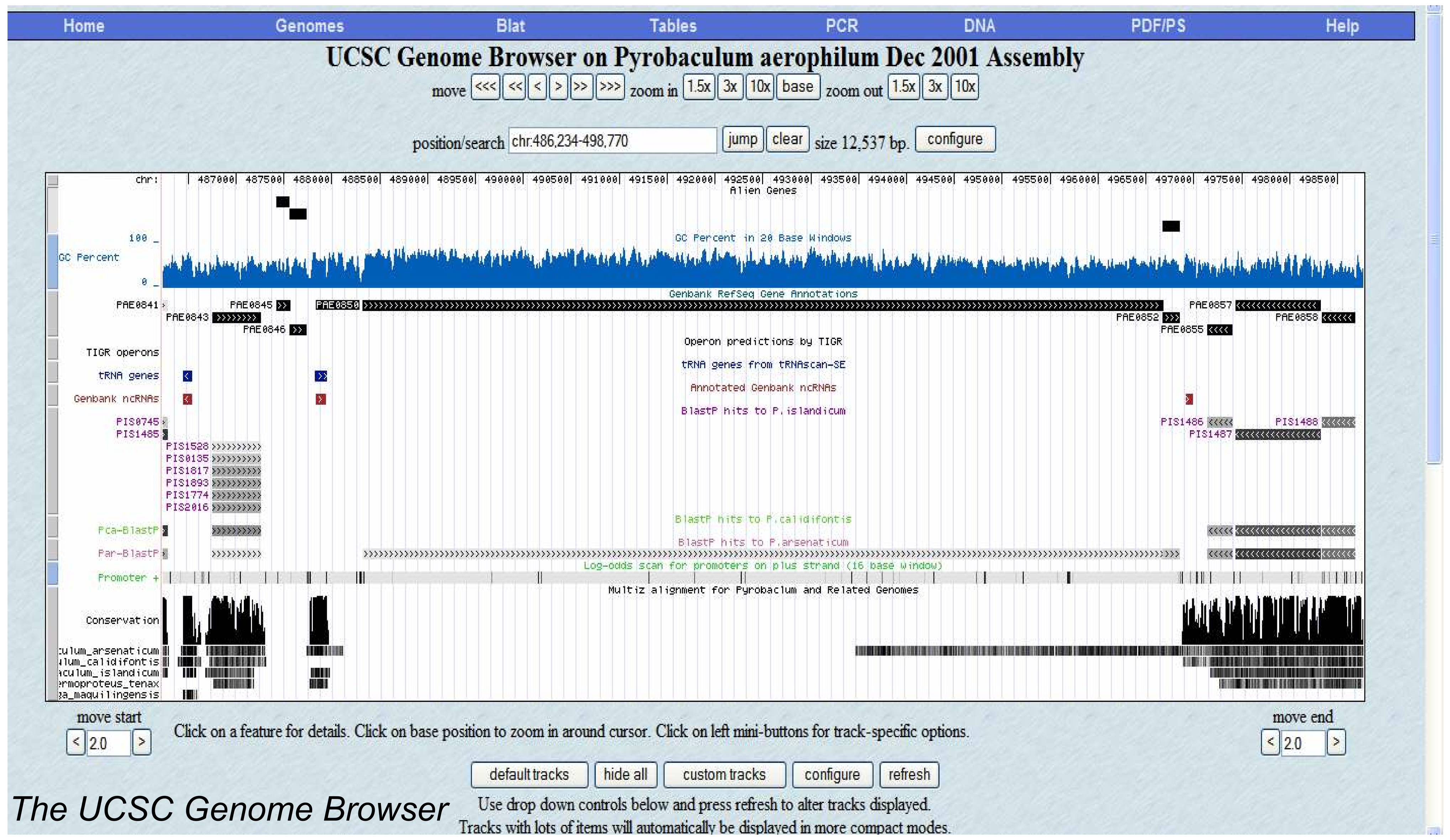
Why is it important?

- Attempting to discover/annotate every important gene by hand would be nearly impossible.
- Early algorithms had lower accuracies, and returned many false positives.
- Better algorithms result in less 'busy' work for the researcher and more quality time investigating the results.



snoRNA Structure

Picture from Todd Lowe



The UCSC Genome Browser

Picture from Todd Lowe

Finding SnoRNAs

- RNA molecules are an intermediate step between DNA and proteins
- However, not all genes code for proteins. Some form RNA molecules that perform other functions
- Small nucleolar RNAs (snoRNAs) modify other non-coding RNAs like rRNAs and tRNAs
- There is some evidence that suggests they can act on mRNAs as well, but this is not well documented

- The original algorithm uses the highly conserved structure of snoRNAs to probabilistically determine potential snoRNAs based on rRNA modifications.
- A basic tenant of comparative genomics is that important sequences are conserved between genomes.
- The new algorithm compares the snoRNAs of different genomes, increasing the overall accuracy.
- This process also allows multiple genomes to be analyzed at once, speeding up the annotation process

Thank you

David Bernick Todd Lowe

This work was completed as part of UCSC's SURF-IT summer undergraduate research program, an NSF CISE REU Site. This material is based upon work supported by the National Science Foundation under Grant No. CCF-0552688

Finding Highly Expressed and 'Alien' Genes

- Not all genes in a genome are equally interesting to a researcher.
- Two interesting categories are highly expressed genes, whose proteins appear in higher amounts in the organism, and 'alien' genes, whose genes appear to be foreign to the genome.
- Highly expressed gene predictions can be used to determine metabolic pathways, while alien gene predictions can be used to examine evolutionary questions.

- The algorithm uses a 'codon usage' analysis. It is postulated that similarly expressed proteins will have similar codon frequencies.
- Each codon is a sequence of three nucleotides that codes for an amino acid.
- If a gene shares similar codon frequencies with known highly expressed genes and shows differences from the rest of the genome, it is marked as "highly expressed."
- If a gene's codon usage differs from both the whole genome AND the known highly expressed genes, then it is marked as an 'alien' gene.

***Pyrococcus horikoshii* sRNAs** *snoRNA results*

Name		C Box		D' Box		C' Box		D Box	
sR1	AAAGAAGGCG	ATGATGA	AGCCTTCCGCAC	CTGA	ATGA	TGAGGA	GTGGACGGCTTC	CTGA	GCCTACTCCT
sR2	AAAAAGAGGG	ATGATGA	GTTTTTCCCTCACT	CTGA	GGAG	TGATGA	GGAGCCGATGCA	CTGA	CCTCGATCAT
sR3	AATTGTGGCG	ATGATGA	ATAGCAAGCCAG	CTGA	AGAG	TGATGA	AGTGAACACCCC	CTGA	GCCTACCTAA
sR4	CCGAGTTGGG	ATGATGA	GGGAGATTTCGG	CCGA	GTGG	TGAGGA	GACTCGCATGGG	CTGA	CCTTTCTAAG
sR5	ATAAGGTGTG	ATGATGA	ACGCCATCGATA	CTGA	GATA	TGATGA	CCGGATTCCTGG	CTGA	TTTCTTTTAT
sR6	TGGAAATGGG	ATGATGA	AGTTTGCTACCC	CTGA	AGAA	TGATGA	ACCCTGCCGTTA	CTGA	CCATTTAGCT
sR7	GTAATCCGGG	ATGATGA	ACCTCCATCCCAA	CTGA	ATAAA	TGATGA	ATGCACATCAGG	CTGA	CATTACCTTT
sR8	CGAATAGGCG	ATGATGA	GCTCCATCCCTAC	CTGA	GTTG	TGATGA	ATGTAGCGCGCT	CTGA	GCAACCCTTA
sR9	CCGACAGGGG	ATGAAGA	GCTTTTGCTTTG	CTGA	GCAGA	TGATGA	CCACGCCCTTCG	CTGA	CCTGCTATTT
sR10	TTAGCGACTA	ATGATGA	ACTACTCCCGG	CTGA	GGGG	TGATGA	ACCACCTACCGG	CTGA	GGTGAAAGCA
sR11	CTCTAGCGGG	ATGATGA	CTTTGCCGAGTG	CTGA	GCTGG	TGATGA	GTAAACAGTCGT	CTGA	CTTTCCCTTT
sR12	ATCGGCTTGG	ATGATGA	GCGTTTACCGGT	CTGA	GCTG	TGATGA	TATCGGCACTGT	CTGA	CTAGTAATCT
sR13	TCATTGTTGG	GTGATGA	GATGGCGGATTG	CTGA	GAGA	TGATGA	GGACCTTAGGGG	CTGA	TTAAATTTCG
sR14	CTATTAACCA	ATGATGA	CGGATCAACCGG	CTGA	TCGAA	TGATGA	CGTCCGCATCAC	CTGA	GGGTTTCAAG
sR15	AGATGAAGAG	ATGATGA	GTAAACCCGTTG	CTGA	GCGG	TGATGA	GAGGATCGACTAG	CTGA	ACAACATCTT
sR16	ATCAAGTCTG	ATGATGA	ACCTTCCCCTCAC	CTGA	AAGG	TGATGA	GCACACCGGTAGG	CTGA	GGGTGATAAT
sR17	ACGGTCTGCG	ATGATGA	GAGCGAACTGCA	CTGA	AAAG	TGATGA	CAGGGCCTTG	CTGA	GCGGTGATCG
sR18	TTTATTTTTA	ATGATGA	AACAGCCAGGACC	CTGA	TGGGA	TGATGA	GTGGTGGGCTTAG	CTGA	TGTTGCGGTA
sR19	TGGCGGGCTC	GTGATGA	GCTTTCCCTACGGC	CCGA	GCTTAGG	CGATGA	GGAATACAGCCAGGG	CTGA	TTTTGGTGAT
sR20	ATGAATGGCG	ATGATGA	GGCCTCGATTGG	CTGA	AT	CGATGA	TTGAGAGGGACTTGG	CTGA	GCGGTGATTA
sR21	TACCATGCCG	ATGATGA	GACCGGTACTGG	CCGA	AGTA	TGATGA	GCACTCGGTTAG	CTGA	GGCCTGAAAA
sR22	GGAACATCCG	ATGATGG	GAACAGGGTAGCTG	CCGA	[illegible]	[illegible]	[illegible]	CTGA	GGAAAGAAAA

Picture from Todd Lowe