Lecture 17

Comparative genomics I:

Genome annotation using evolutionary signatures
Module V: Comparative genomics and evolution

• Today: Whole-genome comparative genomics
  – Evolutionary signatures for systematic genome annotation

• Next week: Phylogenetics and Phylogenomics
  – Distance-based and model-based phylogenetics approaches
  – Gene trees and species trees, reconciliation, coalescence

• Computational foundations:
  – Evolutionary rates and models of evolution
  – Dynamic programming on two-dimensional tree structures
  – Synteny-based alignment, genome assembly
Key goal: Evolution preserves functional elements

We can ‘read’ evolution to reveal functional elements

Yeast (Kellis et al, Nature 2003), Mammals (Xie, Nature 2005), Fly (Stark et al, Nature 07)
Comparative Genomics

Lecture 17 (Today):

Using evolution to study genomes

Evolution

Using genomics to study evolution

Lectures 18-19 (Thursday):
Comparative genomics I: Evolutionary signatures

- **Nucleotide conservation: evolutionary constraint**
  - Purifying selection, neutral branch length, discovery power
  - Detect constrained elements: nucleotides, windows, HMM
  - Estimate fraction constrained: signal vs. background

- **Evolutionary signatures: focus on pattern of change**
  - Different functions ⇔ Characteristic patterns of evolution

- **Signatures of protein-coding genes**
  - Reading-frame conservation, codon-substitution frequency
  - Likelihood ratio framework: Estimating $Q_C Q_N$, scoring
  - Revise genes, read-through, excess constraint regions

- **Signatures of microRNA genes**
  - Structural and evolutionary features of microRNAs
  - Combining features: decision trees, random forests
  - Sense/anti-sense miRNAs, mature/star arm cooperation

- **Measuring selection within the human lineage**
Comparative genomics for genome annotation

- Compare related species to discover functional elements
- Evolution process: random mutation, natural selection
  - Non-functional regions: accumulate mutations, kept
  - Functional regions: accumulate mutations, decrease fitness
  - Evolutionary time: less fit organisms & their genes thin out
Power of many closely related: total branch length

- More branch length $\Rightarrow$ more events $\Rightarrow$ more power
  - Goal: functional vs. non-functional based on # of mutations
  - Very close distances: no mutations in either region
  - Sufficient distance: ability to distinguish increases
  - Very far distances: functional regions no longer conserved

- Many closely related species $\gg$ few distantly related
  - For same total branch length: prefer many close species
  - Functional regions conserved for each pair of species
  - Non-functional regions accumulate noise independently
  - Analogy: recording a concert with multiple microphones

© Source unknown. All rights reserved. This content is excluded from our Creative Commons license. For more information, see http://ocw.mit.edu/help/faq-fair-use/.
Comparative genomics I: Evolutionary signatures

- **Nucleotide conservation: evolutionary constraint**
  - Purifying selection, neutral branch length, discovery power
  - Detect constrained elements: nucleotides, windows, HMM
  - Estimate fraction constrained: signal vs. background

- **Evolutionary signatures: focus on pattern of change**
  - Different functions ⇔ Characteristic patterns of evolution

- **Signatures of protein-coding genes**
  - Reading-frame conservation, codon-substitution frequency
  - Likelihood ratio framework: Estimating $Q_C Q_N$, scoring
  - Revise genes, read-through, excess constraint regions

- **Signatures of microRNA genes**
  - Structural and evolutionary features of microRNAs
  - Combining features: decision trees, random forests
  - Sense/anti-sense miRNAs, mature/star arm cooperation
Genome-wide alignments reveal orthologous segments

- Genome-wide alignments span entire genome
- Comparative identification of functional elements
Comparative genomics and evolutionary signatures

- **Comparative genomics can reveal functional elements**
  - For example: exons are deeply conserved to mouse, chicken, fish
  - Many other elements are also strongly conserved: exons / regulatory?

- **Develop methods for estimating the level of constraint**
  - Count the number of edit operations, number of substitutions and gaps
  - Estimate the number of mutations (including estimate of back-mutations)
  - Incorporate information about neighborhood: conservation ‘windows’
  - Estimate the probability of a constrained ‘hidden state’: HMMs next week
  - Use phylogeny to estimate tree mutation rate, or ‘rejected substitutions’
  - Allow different portions of the tree to have different rates: phylogenetics
Detecting rates and patterns of selection ($\omega/\pi$)

**Estimating intensity of constraint ($\omega$):**
- Probabilistic model of substitution rate
- Maximum Likelihood (ML) estimation of $\omega$
  - Report rate $\omega$
  - Report log odds score that non-neutral
- Window-based vs. sitewise application

**Detect unusual substitution pattern ($\pi$):**
- Probabilistic model of stationary distribution that is different from background.
- ML estimator ($\pi$) of this vector
  - Report PWM for each k-mer in genome.
  - Report log odds score that non-neutral

Manuel Garber, Or Zuk, Xiaohui Xie
Measuring constraint at individual nucleotides

- Reveal individual transcription factor binding sites
- Within motif instances reveal position-specific bias
- More species: motif consensus directly revealed
Detect SNPs that disrupt conserved regulatory motifs

- Functionally-associated SNPs enriched in states, constraint
- Prioritize candidates, increase resolution, disrupted motifs
Comparative genomics I: Evolutionary signatures

- **Nucleotide conservation: evolutionary constraint**
  - Purifying selection, neutral branch length, discovery power
  - Detect constrained elements: nucleotides, windows, HMM
  - Estimate fraction constrained: signal vs. background

- **Evolutionary signatures: focus on pattern of change**
  - Different functions ⇔ Characteristic patterns of evolution

- **Signatures of protein-coding genes**
  - Reading-frame conservation, codon-substitution frequency
  - Likelihood ratio framework: Estimating $Q_C Q_N$, scoring
  - Revise genes, read-through, excess constraint regions

- **Signatures of microRNA genes**
  - Structural and evolutionary features of microRNAs
  - Combining features: decision trees, random forests
  - Sense/anti-sense miRNAs, mature/star arm cooperation
Estimating portion of the genome under constraint

Constraint calculated over a 50mer

Constraint calculated over a 12mer

© Source unknown. All rights reserved. This content is excluded from our Creative Commons license. For more information, see http://ocw.mit.edu/help/faq-fair-use/.
Estimating total fraction under constraint

- Actual distribution of conservation scores (Signal) vs. expected distribution if no constraint (Background).
- At any cutoff: true positives (TP) and false predictions (FP)
- Can’t detect all constrained elements since curves overlap
- But we can estimate the total amount of excess constraint by integrating over entire area between the two curves
Detection of evolutionarily constrained elements

Most new elements in intronic/intergenic regions

Highest enrichment for coding transcripts

Excess positive/purifying selection Distribution of constraint
Coverage depth higher in functional regions

Challenges of low-coverage genomes: varying alignment depth
Evidence of selection against deletions in functional regions
Increase in power from HMRD to 29 mammals

<table>
<thead>
<tr>
<th></th>
<th>$\pi$ log-odds (12mers)</th>
<th>$\pi$ log-odds (50mers)</th>
<th>$\omega$ (12mers)</th>
<th>$\omega$ (50mers)</th>
</tr>
</thead>
<tbody>
<tr>
<td>29 mammals</td>
<td>7.1/1.5/4.6</td>
<td>6.8/1.8/4.1</td>
<td>5.7/1.1/3.8</td>
<td>5.7/1.8/3.0</td>
</tr>
<tr>
<td>(HMRD) Human</td>
<td>4.2/0.0/0.0</td>
<td>5.3/0.1/0.3</td>
<td>4.5/0.0/0.0</td>
<td>5.1/0.6/1.7</td>
</tr>
<tr>
<td>Mouse</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Rat</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Dog</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Estimated / kmers detectable at 5% FDR / base pairs detectable at 5% FDR

Small increase in estimate of genome percentage under constraint
Dramatic increase in power to detect small constrained elements

Manuel Garber, Or Zuk
Comparative genomics I: Evolutionary signatures

- **Nucleotide conservation: evolutionary constraint**
  - Purifying selection, neutral branch length, discovery power
  - Detect constrained elements: nucleotides, windows, HMM
  - Estimate fraction constrained: signal vs. background

- **Evolutionary signatures: focus on pattern of change**
  - Different functions ⇔ Characteristic patterns of evolution

- **Signatures of protein-coding genes**
  - Reading-frame conservation, codon-substitution frequency
  - Likelihood ratio framework: Estimating $Q_CQ_N$, scoring
  - Revise genes, read-through, excess constraint regions

- **Signatures of microRNA genes**
  - Structural and evolutionary features of microRNAs
  - Combining features: decision trees, random forests
  - Sense/anti-sense miRNAs, mature/star arm cooperation
Comparative genomics and evolutionary signatures

- Comparative genomics can reveal functional elements
  - For example: exons are deeply conserved to mouse, chicken, fish
  - Many other elements are also strongly conserved: exons / regulatory?

- Can we also pinpoint specific functions of each region? Yes!
  - Patterns of change distinguish different types of functional elements
  - Specific function ⇔ Selective pressures ⇔ Patterns of mutation/inse/del

- Develop evolutionary signatures characteristic of each function
  Stark et al, Nature 2007
Evolutionary signatures for diverse functions

Protein-coding genes
- Codon Substitution Frequencies
- Reading Frame Conservation

RNA structures
- Compensatory changes
- Silent G-U substitutions

microRNAs
- Shape of conservation profile
- Structural features: loops, pairs
- Relationship with 3’UTR motifs

Regulatory motifs
- Mutations preserve consensus
- Increased Branch Length Score
- Genome-wide conservation

Implications for genome annotation / regulation

- Novel protein-coding genes
- Revised gene annotations
- Unusual gene structures
- Novel structural families
- Targeting, editing, stability
- Riboswitches in mammals
- Novel/expanded miR families
- miR/miR* arm cooperation
- Sense/anti-sense miR switches
- Novel regulatory motifs
- Regulatory motif instances
- TF/miRNA regulatory networks
- Single binding site resolution

Comparative genomics I: Evolutionary signatures

- **Nucleotide conservation: evolutionary constraint**
  - Purifying selection, neutral branch length, discovery power
  - Detect constrained elements: nucleotides, windows, HMM
  - Estimate fraction constrained: signal vs. background

- **Evolutionary signatures: focus on pattern of change**
  - Different functions ↔ Characteristic patterns of evolution

- **Signatures of protein-coding genes**
  - Reading-frame conservation, codon-substitution frequency
  - Likelihood ratio framework: Estimating $Q_C Q_N$, scoring
  - Revise genes, read-through, excess constraint regions

- **Signatures of microRNA genes**
  - Structural and evolutionary features of microRNAs
  - Combining features: decision trees, random forests
  - Sense/anti-sense miRNAs, mature/star arm cooperation
Evolutionary signatures for protein-coding genes

• Same conservation levels, distinct patterns of divergence
  – Gaps are multiples of three (preserve amino acid translation)
  – Mutations are largely 3-periodic (silent codon substitutions)
  – Specific triplets exchanged more frequently (conservative subs.)
  – Conservation boundaries are sharp (pinpoint individual splicing signals)

⇒ Evolutionary signatures of protein-coding selection
Evolutionary signatures of protein-coding genes

DNA insertions and deletions can either insert/remove AAs, or totally mangle the remainder of the protein (frameshift).

Some point mutations to the DNA sequence do not change its protein translation at all.

Natural selection tends to tolerate mutations with little/no effect on the protein.
Protein-coding sequences tolerate distinctive types of change

- synonymous
- conservative
- non-conservative
- frame-shifted
- three stop codons

© Source unknown. All rights reserved. This content is excluded from our Creative Commons license. For more information, see http://ocw.mit.edu/help/faq-fair-use/. 27
Known genes stand out

- Substitution typical of protein-coding regions
- Substitution typical of intergenic regions

© Source unknown. All rights reserved. This content is excluded from our Creative Commons license. For more information, see http://ocw.mit.edu/help/faq-fair-use/.
Comparative genomics I: Evolutionary signatures

- **Nucleotide conservation: evolutionary constraint**
  - Purifying selection, neutral branch length, discovery power
  - Detect constrained elements: nucleotides, windows, HMM
  - Estimate fraction constrained: signal vs. background

- **Evolutionary signatures: focus on pattern of change**
  - Different functions ⇔ Characteristic patterns of evolution

- **Signatures of protein-coding genes**
  - Reading-frame conservation, codon-substitution frequency
  - Likelihood ratio framework: Estimating $Q_C Q_N$, scoring
  - Revise genes, read-through, excess constraint regions

- **Signatures of microRNA genes**
  - Structural and evolutionary features of microRNAs
  - Combining features: decision trees, random forests
  - Sense/anti-sense miRNAs, mature/star arm cooperation
Signature 1: Reading frame conservation

<table>
<thead>
<tr>
<th>RFC</th>
<th>Gene</th>
<th>Intergenic</th>
</tr>
</thead>
<tbody>
<tr>
<td>100%</td>
<td><img src="image1" alt="Gene" /></td>
<td><img src="image2" alt="Intergenic" /></td>
</tr>
<tr>
<td>100%</td>
<td><img src="image3" alt="Gene" /></td>
<td><img src="image4" alt="Intergenic" /></td>
</tr>
<tr>
<td>100%</td>
<td><img src="image5" alt="Gene" /></td>
<td><img src="image6" alt="Intergenic" /></td>
</tr>
<tr>
<td>100%</td>
<td><img src="image7" alt="Gene" /></td>
<td><img src="image8" alt="Intergenic" /></td>
</tr>
<tr>
<td>100%</td>
<td><img src="image9" alt="Gene" /></td>
<td><img src="image10" alt="Intergenic" /></td>
</tr>
<tr>
<td>100%</td>
<td><img src="image11" alt="Gene" /></td>
<td><img src="image12" alt="Intergenic" /></td>
</tr>
<tr>
<td>100%</td>
<td><img src="image13" alt="Gene" /></td>
<td><img src="image14" alt="Intergenic" /></td>
</tr>
<tr>
<td>100%</td>
<td><img src="image15" alt="Gene" /></td>
<td><img src="image16" alt="Intergenic" /></td>
</tr>
<tr>
<td>100%</td>
<td><img src="image17" alt="Gene" /></td>
<td><img src="image18" alt="Intergenic" /></td>
</tr>
</tbody>
</table>

- **Conserved**
- **Mutation**
- **Gap**
- **Frameshift**

**RFC**
- 60%
- 55%
- 90%
- 40%
- 60%
- 100%
- 20%
- 30%
- 40%

**Genes**
- Mutations: 30%
- Gaps: 1.3%
- Frameshifts: 0.14%

**Intergenic**
- Mutations: 58%
- Gaps: 14%
- Frameshifts: 10.2%

**Separation**
- 2-fold
- 10-fold
- 75-fold
## Reading Frame Conservation Test

### Gene

<table>
<thead>
<tr>
<th>100%</th>
<th>100%</th>
<th>100%</th>
<th>100%</th>
<th>100%</th>
<th>100%</th>
</tr>
</thead>
</table>

### Intergenic

| 60% | 60% | 90% | 40% | 60% | 100% |

### Diagram

- **Scer**
  - CTTCTAGATTTTCATCTT-GTCGATGTTCAAACAACGTGTTA-----TCAGAGAAACAGCTCTATGAGAAATCAGCTGATG
- **Spar**
  - TATTCATC-TCTCATCTTCATCAATGTTCAAACAGCGTGTTACAGACACAGAGAAACAGCTTC-TGAGAAGTCAGCCGGTG

### RFC

- F1
- F2
- F3

- 100% Conserved
- 100% Mutation
- 100% Gap
- 56% Frameshift
Revisiting gene content with RFC test

<table>
<thead>
<tr>
<th></th>
<th>Accept</th>
<th>Reject</th>
</tr>
</thead>
<tbody>
<tr>
<td>~4000 named genes</td>
<td>99.9%</td>
<td>0.1%</td>
</tr>
<tr>
<td>~300 intergenic regions</td>
<td>1%</td>
<td>99%</td>
</tr>
<tr>
<td>2000 Hypothetical ORFs</td>
<td>1500</td>
<td>500</td>
</tr>
</tbody>
</table>

High sensitivity and specificity

Example of a rejected ORF

ATG

TAA
Comparative genomics I: Evolutionary signatures

- **Nucleotide conservation: evolutionary constraint**
  - Purifying selection, neutral branch length, discovery power
  - Detect constrained elements: nucleotides, windows, HMM
  - Estimate fraction constrained: signal vs. background

- **Evolutionary signatures: focus on pattern of change**
  - Different functions ⇔ Characteristic patterns of evolution

- **Signatures of protein-coding genes**
  - Reading-frame conservation, codon-substitution frequency
  - Likelihood ratio framework: Estimating $Q_C Q_N$, scoring
  - Revise genes, read-through, excess constraint regions

- **Signatures of microRNA genes**
  - Structural and evolutionary features of microRNAs
  - Combining features: decision trees, random forests
  - Sense/anti-sense miRNAs, mature/star arm cooperation
A method to distinguish these evolutionary signatures should:

• Quantify the distinctiveness of all $64^2$ possible codon substitutions

  • Synonymous: very frequent in protein-coding sequences
  
  • Nonsense: much more frequent in non-coding than coding regions

• Model the phylogenetic relationship among the species

  • Multiple apparent substitutions may be explained by one evolutionary event

• Tolerate uncertainty in the input

  • Unknown ancestral sequences
  
  • Alignment gaps, missing data

• Report the [un]certainty of the result

  • Quantify confidence that given alignment is protein-coding
  
  • Units: p-value, bits, decibans, etc.
Codon evolution can be modeled as a Bayesian network

Each site (codon alignment column) is treated independently.

Given the topology and CPDs, we can simulate evolution of an ancestral sequence.

Additionally given extant (leaf) sequences, the ancestral sequences can be inferred.

For $L$ leaves, CPDs total about $(2L - 2) \cdot 64^2$ parameters.

Conditional probability distribution (CPD) giving, for all codons $a$ & $b$, $Pr(dyak = b | Ancestor = a)$
The Bayes net is parameterized as a continuous-time Markov process:

Each CPD is determined by a rate matrix shared throughout the tree and a branch-specific ‘time’ (branch length):

$$\Pr(\text{child} = b|\text{parent} = a; t) = [e^{Qt}]_{a,b}$$

Intuition: The branch lengths specify how much ‘time’ passed between any two nodes. The rate matrix describes the relative frequencies of codon substitutions per unit branch length. Synonymous substitutions have high rates and nonsense substitutions have low rates.

We can obtain maximum likelihood estimates of $$(2L - 2) + 64^2$$ parameters using EM in training data.
Example nucleotide (4x4) rate & substitution matrices

\[ Q = \begin{pmatrix} -4 & 2 & 1 & 1 \\ 2 & -4 & 1 & 1 \\ 1 & 1 & -4 & 2 \\ 1 & 1 & 2 & -4 \end{pmatrix} \]

\[ \text{Pr}(\text{child} = b | \text{parent} = a; t) = [e^{Qt}]_{a,b} \]

\[ e^{Qt} = \sum_{n=0}^{\infty} \frac{t^n}{n!} Q^n \]

is the solution to the system of differential equations describing the Markov process model of evolution.

MatrixExp[Q * 0] // MatrixForm
\[
\begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}
\]

MatrixExp[Q * 0.001] // MatrixForm // NumberForm[#, 4] &
\[
\begin{pmatrix} 0.996 & 0.001993 & 0.000998 & 0.000998 \\ 0.001993 & 0.996 & 0.000998 & 0.000998 \\ 0.000998 & 0.000998 & 0.996 & 0.001993 \\ 0.000998 & 0.000998 & 0.001993 & 0.996 \end{pmatrix}
\]

MatrixExp[Q * 0.01] // MatrixForm // NumberForm[#, 4] &
\[
\begin{pmatrix} 0.9611 & 0.01932 & 0.009803 & 0.009803 \\ 0.01932 & 0.9611 & 0.009803 & 0.009803 \\ 0.009803 & 0.009803 & 0.9611 & 0.01932 \\ 0.009803 & 0.009803 & 0.01932 & 0.9611 \end{pmatrix}
\]

MatrixExp[Q * 0.1] // MatrixForm // NumberForm[#, 4] &
\[
\begin{pmatrix} 0.692 & 0.1432 & 0.08242 & 0.08242 \\ 0.1432 & 0.692 & 0.08242 & 0.08242 \\ 0.08242 & 0.08242 & 0.692 & 0.1432 \\ 0.08242 & 0.08242 & 0.1432 & 0.692 \end{pmatrix}
\]

MatrixExp[Q * 1.0] // MatrixForm // NumberForm[#, 4] &
\[
\begin{pmatrix} 0.2558 & 0.2533 & 0.2454 & 0.2454 \\ 0.2533 & 0.2558 & 0.2454 & 0.2454 \\ 0.2454 & 0.2454 & 0.2558 & 0.2533 \\ 0.2454 & 0.2454 & 0.2533 & 0.2558 \end{pmatrix}
\]

MatrixExp[Q * 10.0] // MatrixForm // NumberForm[#, 4] &
\[
\begin{pmatrix} 0.25 & 0.25 & 0.25 & 0.25 \\ 0.25 & 0.25 & 0.25 & 0.25 \\ 0.25 & 0.25 & 0.25 & 0.25 \\ 0.25 & 0.25 & 0.25 & 0.25 \end{pmatrix}
\]

Analogy: \( y(t) = e^{qt} \)
solves the differential equation
\[
\frac{dy}{dt} = qy
\]

Side note: Jukes-Cantor and Kimura models are set up so that the entries of \( e^{Qt} \) have closed-form solutions.
The hairy math: how do we estimate $Q$?

- Collect many alignments of known protein-coding sequences (training data)

- Consider the probability of the training data as a function of $Q$

  $$\text{Likelihood}(Q) = \Pr(\text{Training Data}; Q, t)$$

  Still computed using Felsenstein algorithm

- Choose the $Q$ that maximizes that probability:

  $$\hat{Q} = \arg\max_Q (\text{Likelihood}(Q))$$

  Note: $Q$ represents thousands of parameters

- Maximization strategies: expectation-maximization; gradient ascent; simulated annealing; spectral decomposition; others

- Branch lengths can also be optimized in the same way (simultaneously)

- Non-coding model estimated similarly, with random non-coding regions as training data.
Given this generative model of codon evolution:

We can compute the probability of any given alignment, marginalizing over all possible ancestral sequences, using Felsenstein’s pruning algorithm.

If I simulate alignments randomly according to the model, I’ll get this exact alignment once every $10^{117}$ samples.

$$Pr(\text{Leaves}; Q, t) = \frac{1}{10^{117}}$$

$$Pr(\text{Leaves}; Q, t) = \frac{1}{10^{275}}$$
Now suppose we’ve estimated two rate matrices:

\[ Q_C \] estimated from known coding regions

\[ Q_N \] estimated from non-coding regions

These specify different rates of codon substitution, which in turn lead to different probabilities of any given alignment:

\[
Pr(\text{Leaves}; Q_C, t) = \frac{1}{10^{117}}
\]

\[
Pr(\text{Leaves}; Q_N, t) = \frac{1}{10^{152}}
\]
This alignment is $10^{35}$ times more probable under the coding model than the non-coding model.

This alignment is $10^{21}$ times less probable under the coding model than the non-coding model.

This likelihood ratio $\frac{\Pr(\text{Leaves}; Q_C, t)}{\Pr(\text{Leaves}; Q_N, t)}$ is our measure of confidence that the alignment is protein-coding.
Comparative genomics I: Evolutionary signatures

- **Nucleotide conservation: evolutionary constraint**
  - Purifying selection, neutral branch length, discovery power
  - Detect constrained elements: nucleotides, windows, HMM
  - Estimate fraction constrained: signal vs. background

- **Evolutionary signatures: focus on pattern of change**
  - Different functions $\Leftrightarrow$ Characteristic patterns of evolution

- **Signatures of protein-coding genes**
  - Reading-frame conservation, codon-substitution frequency
  - Likelihood ratio framework: Estimating $Q_C Q_N$, scoring
  - Revise genes, read-through, excess constraint regions

- **Signatures of microRNA genes**
  - Structural and evolutionary features of microRNAs
  - Combining features: decision trees, random forests
  - Sense/anti-sense miRNAs, mature/star arm cooperation
Evolutionary signatures can predict new genes and exons

Evolutionary signatures built into a semi-Markov conditional random field to predict protein-coding exons

Courtesy of Macmillan Publishers Limited. Used with permission.
New protein-coding genes

New genes supported by Illumina BodyAtlas transcripts
Submitted to GENCODE for validation / manual curation
Translational read-through in flies and mammals

One of four novel candidates in the human genome: OPRL1 neurotransmitter

- **New mechanism of post-transcriptional regulation?**
  - Conserved in both mammals (4 candidates) and flies (350 candidates)
  - Strongly enriched for neurotransmitters, brain-expressed proteins, TF regulators
  - After correcting for gene length: TF enrichment remains

- **Evidence suggestive of regulatory control**
  - Read-through stop codon perfectly conserved in 93% of cases (24% at bkgnd)
  - Upstream bases show increased conservation. Downstream is TGAC.
  - GCA triplet repeats
  - Increased RNA secondary structure

Lin *et al.*, Genome Research 2007
Jungreis *et al.*, in preparation
Discover of translational readthrough genes

Discovery of 4 readthrough genes, abundant in many animal genomes
Overlapping selection in protein-coding exons

Overlapping synonymous constrained elements

Roles in splicing, translation, regulation

© Source unknown. All rights reserved. This content is excluded from our Creative Commons license. For more information, see http://ocw.mit.edu/help/faq-fair-use/.
Codon-specific measures of positive selection

Gene-wide vs. punctate regions of exons positive selection
Comparative genomics I: Evolutionary signatures

- **Nucleotide conservation: evolutionary constraint**
  - Purifying selection, neutral branch length, discovery power
  - Detect constrained elements: nucleotides, windows, HMM
  - Estimate fraction constrained: signal vs. background

- **Evolutionary signatures: focus on pattern of change**
  - Different functions ⇔ Characteristic patterns of evolution

- **Signatures of protein-coding genes**
  - Reading-frame conservation, codon-substitution frequency
  - Likelihood ratio framework: Estimating $Q_CQ_N$, scoring
  - Revise genes, read-through, excess constraint regions

- **Signatures of microRNA genes**
  - Structural and evolutionary features of microRNAs
  - Combining features: decision trees, random forests
  - Sense/anti-sense miRNAs, mature/star arm cooperation
New RNA structures and families

<table>
<thead>
<tr>
<th></th>
<th>No. of structures</th>
<th>No. of novel structures</th>
<th>No. of families</th>
<th>No. of novel families</th>
<th>EvoFold score</th>
<th>RNAz overlap enrichment (x)</th>
<th>DNAse hypersensitivity overlap (%)</th>
<th>Avg. correlation of tissue-specific expression within families</th>
</tr>
</thead>
<tbody>
<tr>
<td>EvoFold all (no CDS)</td>
<td>27,012</td>
<td>26,643</td>
<td>n/a</td>
<td>n/a</td>
<td>14</td>
<td>13.5</td>
<td>25 (P ≤ 5e−3)</td>
<td>n/a</td>
</tr>
<tr>
<td>Unfiltered families</td>
<td>3293</td>
<td>3081</td>
<td>1254</td>
<td>1215</td>
<td>18</td>
<td>17.3</td>
<td>25 (P ≤ 7e−3)</td>
<td>0.14 (P ≤ 1e−3)</td>
</tr>
<tr>
<td>Filtered families</td>
<td>725</td>
<td>526</td>
<td>220</td>
<td>181</td>
<td>18</td>
<td>29</td>
<td>32 (P ≤ 4e−3)</td>
<td>0.15 (P ≤ 1e−3)</td>
</tr>
</tbody>
</table>

New structs fall in families, supported by evolution/energy

Ex: new struct in XIST long non-coding RNA

Known function in X-chromosome inactivation

Possible functional domain of XIST?
RNA families: orthologous/paralogous conservation

Example of new structural 3’UTR family in MAT2A gene likely role in detecting S-adeosyl-methionic (SAM) level
Computational challenge of miRNA discovery

760,355 miRNA-like hairpins

60-100 true miRNAs

A false positive rate of 0.5% $\Rightarrow$ 3800 spurious hairpins.
Need 99.99% specificity (>5,000-fold enrichment)
Evolutionary signatures for microRNA genes

miRNAs show characteristic conservation properties.

© Source unknown. All rights reserved. This content is excluded from our Creative Commons license. For more information, see http://ocw.mit.edu/help/faq-fair-use/.
Distinguishing true miRNAs from random hairpins

Evolutionary features

(1) Correlation with conservation profile

(2) MFE of the consensus fold

(3) Structure conservation index

Structural features

(4) Hairpin stability (MFE z-score)

(5) Number of asymmetric loops

(6) Number of symmetric loops

Feature performance

<table>
<thead>
<tr>
<th>Feature</th>
<th>Enrichment</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total</td>
<td>4551</td>
</tr>
<tr>
<td>Cons. Profile</td>
<td>327</td>
</tr>
<tr>
<td>Arm / Loop Cons.</td>
<td>19</td>
</tr>
<tr>
<td>Arm Cons.</td>
<td>5</td>
</tr>
<tr>
<td>Structure Cons.</td>
<td>6</td>
</tr>
<tr>
<td>Hairpin Energy</td>
<td>39</td>
</tr>
<tr>
<td>% Paired Bases</td>
<td>2.3</td>
</tr>
<tr>
<td>Arm length</td>
<td>1.7</td>
</tr>
<tr>
<td>Loop Length</td>
<td>1.4</td>
</tr>
<tr>
<td>Structure Length</td>
<td>1.4</td>
</tr>
<tr>
<td>Loop Symmetry</td>
<td>3</td>
</tr>
</tbody>
</table>

Combination of features: > 4,500-fold enrichment
Comparative genomics I: Evolutionary signatures

• Nucleotide conservation: evolutionary constraint
  – Purifying selection, neutral branch length, discovery power
  – Detect constrained elements: nucleotides, windows, HMM
  – Estimate fraction constrained: signal vs. background

• Evolutionary signatures: focus on pattern of change
  – Different functions ⇔ Characteristic patterns of evolution

• Signatures of protein-coding genes
  – Reading-frame conservation, codon-substitution frequency
  – Likelihood ratio framework: Estimating $Q_C Q_N$, scoring
  – Revise genes, read-through, excess constraint regions

• Signatures of microRNA genes
  – Structural and evolutionary features of microRNAs
  – Combining features: decision trees, random forests
  – Sense/anti-sense miRNAs, mature/star arm cooperation
miRNA detection using many decision trees

- For each tree:
  - Randomly select:
    - Subset of features to base classification on
    - Subset of +/- training examples
    - Remainder of testing examples
  - Use to train a decision tree classifier:
    - Select a feature and cutoff at each level
    - Continue with feature/cutoff at next level
    - (...)
  - Evaluate performance on test set:
    - Push each element down the decision tree
    - Leaf label gives classification decision

- To combine trees:
  - Average prediction class across trees
  - Report class with maximum # of votes
Random Forests: Combine many decision trees

- **Many decision trees:**
  - Each can select cutoffs and direction of cutoff
  - Each feature can be reused multiple times
  - Used serially (AND) and in parallel (OR)

- **Ensemble classifier**
  - Bagging: model averaging, combines predictions
  - Can take median of predictions

- **Advantages:** Robustness, Feature importance
Evidence 1: Novel miRNAs match sequencing reads

Ruby, Bartel, Lai

348 reads
16 reads

© Source unknown. All rights reserved. This content is excluded from our Creative Commons license. For more information, see http://ocw.mit.edu/help/faq-fair-use/.
Evidence 2: Genomic properties typical of miRNAs

- Novel miRNAs in introns of known genes
- Preference for + strand, transcription factors

- Genomic clustering with novel / known miRNAs
- Same family, common origin / same precursor

© Source unknown. All rights reserved. This content is excluded from our Creative Commons license. For more information, see http://ocw.mit.edu/help/faq-fair-use/.
Two ‘dubious’ protein-coding genes are in fact miRNAs

Two novel miRNAs overlap exons (5’UTR and coding!)

- Both CG31044 and CG33311 were independently rejected as *dubious* based on their non-protein-coding conservation patterns (Lin *et al.*)
- Novel miRNA genes provide explanation for their transcripts, as their precursor miRNA

© Source unknown. All rights reserved. This content is excluded from our Creative Commons license. For more information, see [http://ocw.mit.edu/help/faq-fair-use/](http://ocw.mit.edu/help/faq-fair-use/).
Comparative genomics I: Evolutionary signatures

• **Nucleotide conservation: evolutionary constraint**
  – Purifying selection, neutral branch length, discovery power
  – Detect constrained elements: nucleotides, windows, HMM
  – Estimate fraction constrained: signal vs. background

• **Evolutionary signatures: focus on pattern of change**
  – Different functions ⇔ Characteristic patterns of evolution

• **Signatures of protein-coding genes**
  – Reading-frame conservation, codon-substitution frequency
  – Likelihood ratio framework: Estimating $Q_C Q_N$, scoring
  – Revise genes, read-through, excess constraint regions

• **Signatures of microRNA genes**
  – Structural and evolutionary features of microRNAs
  – Combining features: decision trees, random forests
  – Sense/anti-sense miRNAs, mature/star arm cooperation
Surprise 1: microRNA & microRNA* function

- Both hairpin arms of a microRNA can be functional
  - High scores, abundant processing, conserved targets
  - Hox miRNAs miR-10 and miR-iab-4 as master Hox regulators

Stark et al, Genome Research 2007
Evidence of miR-iab-4 anti-sense (AS) function

• A single miRNA locus transcribed from both strands
• The two transcripts show distinct expression domains (mutually exclusive)
• Both processed to mature miRNAs: mir-iab-4, miR-iab-4AS (anti-sense)

© Cold Spring Harbor Laboratory Press. All rights reserved. This content is excluded from our Creative Commons license. For more information, see http://ocw.mit.edu/help/faq-fair-use/.
miR-iab-4AS leads to homeotic transformations

- Mis-expression of mir-iab-4S & AS: alters \(\rightarrow\) wings homeotic transform.
- Stronger phenotype for AS miRNA
- Sense/anti-sense pairs as general building blocks for miRNA regulation
- 10 sense/anti-sense miRNAs in mouse

© Cold Spring Harbor Laboratory Press. All rights reserved. This content is excluded from our Creative Commons license. For more information, see [http://ocw.mit.edu/help/faq-fair-use/](http://ocw.mit.edu/help/faq-fair-use/).

Comparative genomics I: Evolutionary signatures

- **Nucleotide conservation: evolutionary constraint**
  - Purifying selection, neutral branch length, discovery power
  - Detect constrained elements: nucleotides, windows, HMM
  - Estimate fraction constrained: signal vs. background

- **Evolutionary signatures: focus on pattern of change**
  - Different functions ⇔ Characteristic patterns of evolution

- **Signatures of protein-coding genes**
  - Reading-frame conservation, codon-substitution frequency
  - Likelihood ratio framework: Estimating $Q_C$, $Q_N$, scoring
  - Revise genes, read-through, excess constraint regions

- **Signatures of microRNA genes**
  - Structural and evolutionary features of microRNAs
  - Combining features: decision trees, random forests
  - Sense/anti-sense miRNAs, mature/star arm cooperation

- **Measuring selection within the human lineage**
Mammalian constraint matches Human SNPs

Genome-wide agreement of selection and polymorphisms
Human constraint outside conserved regions

- **Non-conserved regions:**
  - ENCODE-active regions show reduced diversity
  - Lineage-specific constraint in biochemically-active regions

- **Conserved regions:**
  - Non-ENCODE regions show increased diversity
  - Loss of constraint in human when biochemically-inactive

Average diversity (heterozygosity)
Aggregate over the genome
Strongest: motifs, short RNA, Dnase, ChIP, IncRNA

• Significant derived allele depletion in active features
Bound motifs show increased human constraint

Position-specific reduction in bound motif heterozygosity

Aggregate across thousands of CTCF motif instances
Most constrained human-specific enhancer functions

- Transcription initiation from Pol2 promoter
- Transcription coactivator activity
- Transcription factor binding
- Chromatin binding
- Negative regulation of transcription, DNA-dependent
- Transcription factor complex
- Protein complex
- Protein kinase activity
- Nerve growth factor receptor signaling pathway
- Signal transducer activity
- Protein serine/threonine kinase activity
- Negative regulation of transcription from Pol2 promoter
- Protein tyrosine kinase activity
- In utero embryonic development

© Source unknown. All rights reserved. This content is excluded from our Creative Commons license. For more information, see http://ocw.mit.edu/help/faq-fair-use/.

Regulatory genes: Transcription, Chromatin, Signaling.
Developmental enhancers: embryo, nerve growth
Comparative genomics I: Evolutionary signatures

- **Nucleotide conservation: evolutionary constraint**
  - Purifying selection, neutral branch length, discovery power
  - Detect constrained elements: nucleotides, windows, HMM
  - Estimate fraction constrained: signal vs. background

- **Evolutionary signatures: focus on pattern of change**
  - Different functions $\Leftrightarrow$ Characteristic patterns of evolution

- **Signatures of protein-coding genes**
  - Reading-frame conservation, codon-substitution frequency
  - Likelihood ratio framework: Estimating $Q_C Q_N$, scoring
  - Revise genes, read-through, excess constraint regions

- **Signatures of microRNA genes**
  - Structural and evolutionary features of microRNAs
  - Combining features: decision trees, random forests
  - Sense/anti-sense miRNAs, mature/star arm cooperation

- **Measuring selection within the human lineage**