Lecture 4
Comparative genomics II:
- Measuring constraint
- Evolutionary signatures

Module 1: Comparative genomics

• Foundations vs. frontiers
  – Foundations: Classical computational methods / biological topics
  – Frontiers: Latest developments, open questions, research areas
  – Duality for each: basic problems / fundamental techniques
• Sequence alignment:
  – Local/global alignment: infer nucleotide-level evolutionary events
  – Database search: scan for regions that may have common ancestry
• Comparative genomics
  – Detect evidence of purifying selection / function: 5% of human
  – Detect specific types of purifying selection: evolutionary signatures

Comparative genomics I: Evolutionary signatures

• Nucleotide conservation vs. evolutionary signatures
• Protein-coding genes
  – Reading-frame conservation, codon-substitution frequency
  – Likelihood ratio framework: Estimating $Q_C$, $Q_N$, scoring
  – Read-through genes, excess constraint regions
• microRNA genes
  – Structural and evolutionary features of microRNAs
  – Combining features: decision trees, random forests
  – Sense/anti-sense miRNAs, mature/star arm cooperation
• Regulatory motifs and motif instances
  – Genome-wide conservation criteria: intergenic, coding, u/d
  – Branch-length score (BLS), control motifs, confidence
  – Inferring regulatory networks and individual binding sites

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

Stark et al, Nature 2007

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

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

This likelihood ratio is our measure of confidence that the alignment is protein-coding.

Comparative genomics I: Evolutionary signatures

- Nucleotide conservation vs. evolutionary signatures
- Protein-coding genes
  - Reading-frame conservation, codon-substitution frequency
  - Likelihood ratio framework: Estimating $Q_C/Q_N$, scoring
    - Read-through genes, excess constraint regions
- microRNA genes
  - Structural and evolutionary features of microRNAs
  - Combining features: decision trees, random forests
  - Sense/anti-sense miRNAs, mature/star arm cooperation
- Regulatory motifs and motif instances
  - Genome-wide conservation criteria: intergenic, coding, u/d
  - Branch-length score (BLS), control motifs, confidence
  - Inferring regulatory networks and individual binding sites

2d. Stop-codon read-through

Evolutionary signatures clearly reflect translation termination

3' ends of some typical genes
An unusual gene (*kelch*)

- Examination of the function of two *kelch* proteins generated by stop codon suppression
- Douglas N. Richman and Lynn Chacon

**SUMMARY**

- *kelch* gene is transcriptionally regulated by a light-dependent mechanism
- Protein-coding
- Continued protein-coding
- No more stop codon

Translational read-through in flies and mammals

- One of four novel candidates in the human genome: OPRL1 neurotransmitter
- Protein-coding conservation
- Continued protein-coding conservation
- No more stop codon

- 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 background)
  - Upstream bases show increased conservation. Downstream is TGAC.
  - GCA triplet repeats
  - Increased RNA secondary structure

A new example (*Caki*)

- No RT
- Single RT
- Double RT

Frame-0 excess argues for read-through

- Alt-splicing or independent ORFs should show no bias

Biased stop-codon usage reveals 'leaky' stops

- TGAC is most frequent in RT, least in genome
- Known to be a 'leaky' stop codon context
As a group, supported by single species evidence

- Z-curve measures codon usage patterns in single species
- The read-through region matches distribution before regular stops
- After 2nd stop matches regions after regular stops

How common is stop codon readthrough in Drosophila?

- Literature: 3

CSF score distribution for all FlyBase genes that have an in-frame ORF of at least 10aa immediately downstream of the annotated stop codon.

Translational read-through in mammals

Four candidates found, mostly neuronal proteins in the adult brain

A look at FOXP2 – Possible 3'UTR function? (not in fish, yes in frog)

Evidence that read-through rate may be regulated

- Unusual stop codon context TGAC
- Increased conservation around stop
- Evolutionarily conserved hairpins
- Increased frequency of GCA repeats

Protein-coding evolutionary signatures in mammals

2e. Excess constraint genes
**Excess constraint in protein-coding sequences**

- **Goal**: identify protein-coding sequences under selection for additional, overlapping functions
- **Approach**: screen all known genes for regions with far fewer-than-expected synonymous substitutions
- The HOXB5 example is obvious and easily detectable with simple methods. But it’s not always so straightforward...

**Protein-coding sequences can simultaneously encode other functional elements**

The 5' end of the HOXB5 ORF appears to encode both amino acids and RNA secondary structure. Such regions should show much less tolerance of synonymous codon substitutions.

**Detecting genes encoding overlapping functional elements**

Across 29 mammals, average codon site shows five synonymous substitutions.

<table>
<thead>
<tr>
<th>Typical windows</th>
<th>Reduced synonymous rates</th>
</tr>
</thead>
<tbody>
<tr>
<td>CFTR</td>
<td>ALDH2</td>
</tr>
<tr>
<td>BMP4</td>
<td>GRIA2</td>
</tr>
<tr>
<td>HOXA2</td>
<td></td>
</tr>
</tbody>
</table>

**Measuring synonymous rates using phylogenetic codon models**

Null model:

- Rate matrix (4x4) Tree
- Rate matrix (64x64) Tree

Alternative:

- \( \omega \) × tree scaled by free parameter \( \omega \)
- synonymous rates scaled by \( \lambda_s \)
- non-synonymous rates scaled by \( \lambda_n \)

- Estimate null model based on lots of data
- Alternative model has only two additional parameters \((\lambda_s, \lambda_n)\), so can be estimated for short windows

**Modeling excess constraint in phylogenomic setting**

- Substitution matrix \( \Phi \), \( \lambda_s \), \( \lambda_n \) scaling factors for syn/nons rates
  - E.g. \( \lambda_s = 0.5 \) means estimates non-syn. rate is \( \frac{1}{2} \) of null model
  - Null model obtained from (a) genome (b) gene
- Evaluate statistical significance of rate estimates.
  - Accounts for alignment coverage, codon degeneracy etc

**Likely role of SCEs in alternative splicing**

Lin et al., in preparation

- SNP
- Insertion
- Deletion
- Alternative splicing
Most Hox genes show overlapping constraint regions

- Example: First 50 amino-acids of HoxB5

Lin et al., in preparation

Example: Hoxa2 overlapping regulatory module

- Rhombomere 2 expression (Tümpel et al., PNAS 2008)
- Rhombomere 4 expression (Lampe et al., NAR 2008)

Example: ADAR splice variant

Dual-coding region in THRA/NR1D1

Example: BRCA1 alternative translation start

Hurst & Pal (2001)

"We find a repeatable pronounced peak...caused by a plummet in the silent-site rate of evolution. The most parsimonious interpretation of these data is that purifying selection is acting on silent sites."


The Trouble with Sliding Windows and the Selective Pressure in BRCA1

"We demonstrate that a previous finding that a particular region of the BRCA1 gene experienced a synonymous rate reduction driven by purifying selection is likely an artifact of the sliding window analysis."

Example: Hoxa2 overlapping regulatory module

- Hindbrain enhancer

Example: Hoxa2 overlapping regulatory module

- Reporter

Example: Hoxa2 overlapping regulatory module

- Reporter

Example: Hoxa2 overlapping regulatory module

- Reporter

Example: Hoxa2 overlapping regulatory module

- Reporter
**Comparative genomics I: Evolutionary signatures**

- **Nucleotide conservation vs. evolutionary signatures**
- **Protein-coding genes**
  - Reading-frame conservation, codon-substitution frequency
  - Likelihood ratio framework: Estimating $Q_cQ_v$, scoring
  - Read-through genes, excess constraint regions
- **microRNA genes**
  - Structural and evolutionary features of microRNAs
  - Combining features: decision trees, random forests
  - Sense/anti-sense miRNAs, mature/star arm cooperation
- **Regulatory motifs and motif instances**
  - Genome-wide conservation criteria: intergenic, coding, u/d
  - Branch-length score (BLS), control motifs, confidence
  - Inferring regulatory networks and individual binding sites

**Remainder of this talk:**

1. Can we define evolutionary signatures for microRNA genes and their targets?
2. Can we use these to gain new insights on microRNA biology, functions, and networks?

**Challenge:** Very subtle signals

**Computational challenge of miRNA discovery**

- 760,355 miRNA-like hairpins
- 60-100 true miRNAs

A false positive rate of 0.5% ➔ 3800 spurious hairpins.
Need 99.99% specificity (>5,000-fold enrichment)

**Evolutionary signatures for microRNA genes**

miRNAs show characteristic conservation properties
Distinguishing true miRNAs from random hairpins

**Evolutionary features**
1. 
2. 
3. 

**Structural features**
4. 
5. 

**Combination of features:** > 4,500-fold enrichment

---

Genome-wide prediction results

- Across fly genome: 101 hairpins above 0.95 cutoff
  - 60 of 74 (81%) known miRNAs rediscovered
  - 24 novel miRNAs discovered and expression-validated (Bartel)
  - 17 additional candidates show diverse evidence of function

---

Evolutionary signatures for mature miRNAs

- Computationally identify exact position of cleavage
- Use correlation with target motifs in 3'UTRs
  - 7-mer complement is avoided in anti-targets
  - Highly conserved in targets
  - Avoided in anti-targets

- Combined score using structural / 3'UTR features
- Rfam miRNAs: 85% within 1bp, and 78% exact

---

Evidence 1: Novel miRNAs match sequencing reads

- Ruby, Bartel, Lai
  - 348 reads
  - 16 reads

---

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

---

Evidence 3: Exclusion of exons / repeats
Two 'dubious' protein-coding genes are in fact miRNAs

- 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

Correction of mature annotation for known miRNAs

- 5' end score peaks at true start (structural and 3'UTR features)
- In 6 cases, discrepancy found, and verified by sequencing
- Leads to drastic changes in target spectrum

Conclusion: Novel miRNAs, families, and targets

- 41 novel miRNA genes (30% increase)
- 37 novel miRNA families (70% increase)
- 500 new target genes (only 15% increase)
- New miRNAs regulate many existing miRNA targets

Increased potential for combinatorial regulation

Comparative genomics I: Evolutionary signatures

- Nucleotide conservation vs. evolutionary signatures
- Protein-coding genes
  - Reading-frame conservation, codon-substitution frequency
  - Likelihood ratio framework: Estimating $Q_{EM}$, scoring
  - Read-through genes, excess constraint regions
- microRNA genes
  - Structural and evolutionary features of microRNAs
  - Combining features: decision trees, random forests
  - Sense/anti-sense miRNAs, mature/star arm cooperation
- Regulatory motifs and motif instances
  - Genome-wide conservation criteria: intergenic, coding, u/d
  - Branch-length score (BLS), control motifs, confidence
  - Inferring regulatory networks and individual binding sites

Three surprises

1. Sense and anti-sense miRNAs may be functional
   - Evolutionary evidence
   - Sequencing evidence

2. Multiple mature miRNAs surround primary product
   - Evolutionary evidence
   - Sequencing evidence

3. Both miRNA and miRNA* show functional targets
   - Evolutionary evidence
   - Sequencing evidence

Evidence of selection and in vivo expression
**Surprise 1: Both strands may be expressed and functional**

- Both strands show processed mature product
- Both strands show precise and distinct embryonic domains

**miR-iab-4 expression**

- Sense Anti-sense

**miRNA hairpin score**

- (+) Hairpin score (+)
- (-) Hairpin score (-)

**0.95 cutoff**

- Both strands score > 0.95
- Both strands show processed mature product
- Both strands show precise and distinct targets

**176 reads**

**Surprise 2: Multiple 5’-ends for a single miRNA arm**

- Some miRNAs show imprecise start
  - Multiple 7-mers show MCS, avoidance
  - Structural features high at multiple starts
    -> Incorrect start site predicted
- These show imprecise processing
  - Sequencing results show 3-fold increased off-peak processing
- Multiple mature products, with potentially functional targets

**Surprise 3: miRNA* scores highly ⇒ highly expressed**

- 13 star arms score very highly
  - In 4 cases, more highly than mature
  - Relationship with 3’UTRs, targets
- 4 show increased star processing
  - 5-fold increase in star product
  - In vivo expression correlated

**Expressed + functional targets?**

**Surprise 4: miR-10/miR-10* as a master Hox regulator**

- miR-10* is functional
  - miR-10* scores highly
  - miR-10* highly expressed
  - miR-10* has many targets
  -> Arm formerly known as miR-10*
- Both arms are functional
  - Each arm has highly conserved target sites in Hox genes
  - Unique: mature/star cooperation
  - Unique: multiple Hox targets
  -> A master regulator for Hox?

**Novel miRNAs validated by sequencing reads**

- In fly genome: 101 hairpins above 0.95 cutoff
- 60 of 74 (81%) known Rfam miRNAs rediscovered
  + 24 novel expression-validated by 454&Solexa (Bartel/Hannon)
- 17 additional candidates show diverse evidence of function

Rely on reads for discovery, use evolutionary signal to study function
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)

miR-iab-4AS leads to homeotic transformations

- Mis-expression of mir-iab-4S & AS: alters 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

MicroRNAs signatures reveal complex Hox network

- Illustrates miR/miR* and miR/miR-AS cooperation

Comparative genomics I: Evolutionary signatures

- Nucleotide conservation vs. evolutionary signatures
- Protein-coding genes
  - Reading-frame conservation, codon-substitution frequency
  - Likelihood ratio framework: Estimating QcQc, scoring
  - Read-through genes, excess constraint regions
- microRNA genes
  - Structural and evolutionary features of microRNAs
  - Combining features: decision trees, random forests
  - Sense/anti-sense miRNAs, mature/star arm cooperation
- Regulatory motifs and motif instances
  - Genome-wide conservation criteria: intergenic, coding, u/d
  - Branch-length score (BLS), control motifs, confidence
  - Inferring regulatory networks and individual binding sites

Evolutionary signatures for regulatory motifs

- Individual motif instances are preferentially conserved
- Measure conservation across entire genome
  - Over thousands of motif instances → increased discovery power
  - Couple to rapid enumeration and rapid string search
- De novo discovery of regulatory motifs

4. Genome-wide discovery of regulatory motifs
Conservation islands overlap known motifs

Test 1: Intergenic conservation

Test 2: Intergenic vs. Coding

Test 3: Upstream vs. Downstream

Genome-wide conservation

Evaluate conservation within:
1. All intergenic regions 13% 2%
2. Intergenic: coding 13% : 3% 2% : 7%
3. Upstream: downstream 12:0 1:1

A signature for regulatory motifs

Motif discovery pipeline

1. Enumerate motif seeds
   TGC — gap — TAG
   • Six non-degenerate characters with variable size gap in the middle
2. Score seed motifs
   • Use a conservation ratio corrected for composition and small counts to rank seed motifs
3. Expand seed motifs
   SRTGCY — gap — WTAGR
   • Use expanded nucleotide IUPAC alphabet to fill unspecified bases around seed using hill climbing
4. Cluster to remove redundancy
   • Using sequence similarity
**Consensus MCS Matches to known Expression enrichment Promoters Enhancers**

1. CTAATTAAA 65.6 engrailed (en) 25.4
2. TTKCAATTAA 57.3 reversed-polarity (repo) 5.8 4.2
3. WATTRATTK 54.9 araucan (ara) 11.7 2.6
4. AAATTTATGCK 54.4 paired (prd) 4.5 16.5
5. GCAATAAA 51 ventral veins lacking (vvl) 13.2 0.3
6. DTAATTTRYNR 46.7 Ultrabithorax (Ubx) 16 3.3
7. TGATTAAT 45.7 apterous (ap) 7.1 1.7
8. YMATTAAAA 43.1 abdominal A (abd-A) 7 2.2
9. AAACNNGTT 41.2
10. RATTKAATT 40 3.9 0.7
11. GCACGTGT 39.5 fushi tarazu (ftz) 17.9
12. AACASCTG 38.8 broad-Z3 (br-Z3) 10.7
13. AATTRMATTA 38.2 19.5 1.2
14. TATGCWAAT 37.8 5.8 2
15. TAATTATG 37.5 Antennapedia (Antp) 14.1 5.4
16. CATNAATCA 36.9 1.8 1.7
17. TTACATAA 36.9 5.4
18. RTAAATCAA 36.3 3.2 2.8
19. AATKNMATTT 36 3.6 0
20. ATGTCAAHT 35.6 2.4 4.6
21. ATAAAYAAA 35.5 57.2 -0.5
22. YYAATCAAA 33.9 5.3 0.6
23. WTTTTATG 33.8 Abdominal B (Abd-B) 6.3 6
24. TTTYMATTA 33.6 extradenticle (exd) 6.7 1.7
25. TGTMAATA 33.2 8.9 1.6
26. TAAYGAG 33.1 4.7 2.7
27. AAAKTGA 32.9 7.6 0.3
28. AAANNAAA 32.9 449.7 0.8
29. RTAAWTTAT 32.9 gooseberry-neuro (gsb-n) 11 0.8
30. TTATTTAYR 32.9 Deformed (Dfd) 30.7

**Power of evolutionary signatures for motif discovery**

- Consensus MCS Matches to known Expression enrichment Promoters Enhancers
- Ability to discover full dictionary of regulatory motifs de novo

**Tissue-specific enrichment and clustering**

- Infer candidate functions for novel motifs
- Reveal ‘modules’ of co-operating motifs

**Discovered motifs show positional biases**

- May represent new core promoter elements
- Show enrichment in distinct functional categories

**Recognizing functional motifs in coding regions**

- Each motif type has distinct signatures
  - DNA: strand symmetric
  - RNA: strand-specific, frame-invariant
  - Protein: strand-specific, frame-biased
- Use frame invariance as a signature
  - Evaluate each frame offset separately
  - Motifs due to di-codon usage biases
  - Conserved in only one frame offset
  - Motifs due to RNA-level regulation
  - Conserved in all three frame offsets
- Result: miRNA motifs in coding exons
  - Top 20 motifs → 11 miRNA seeds (before: 11 seeds in 200+ motifs)
  - Conclude: miRNA targets in coding regions!

**miRNA targeting in protein-coding regions**

- MicroRNA seeds are specifically selected
- Coding & 3’UTRs show same conservation profile

**Comparative genomics I: Evolutionary signatures**

- Nucleotide conservation vs. evolutionary signatures
- Protein-coding genes
  - Reading-frame conservation, codon-substitution frequency
  - Likelihood ratio framework: Estimating QcQm, scoring
  - Read-through genes, excess constraint regions
- microRNA genes
  - Structural and evolutionary features of microRNAs
  - Combining features: decision trees, random forests
  - Sense/anti-sense miRNAs, mature/star arm cooperation
- Regulatory motifs and motif instances
  - Genome-wide conservation criteria: intergenic, coding, u/d
  - Branch-length score (BLS), control motifs, confidence
  - Inferring regulatory networks and individual binding sites
5. Signatures for individual motif instances

### 5. Evolutionary signatures of motif instances

- **Allow for motif movements**
  - Sequencing/alignment errors
  - Loss, movement, divergence

- **Measure branch-length score**
  - Sum evidence along branches
  - Close species little contribution

#### Computing Branch Length Score (BLS)

- **Branch Length Score ➔ Confidence**
  1. Evaluate non-motif probability of a given score
     - Sequence could also be conserved due to overlap with un-annotated element (e.g. non-coding RNA)
  2. Account for differences in motif composition and length
     - For example, short motif more likely to be conserved by chance

#### Control motifs

- Control motifs are the basis of our estimation of the background level of conservation and for evaluating enrichment
- Each motif has its own set of controls
- They are chosen to:
  - Have the same composition as the original motif
  - Match the target regions (e.g. promoters) with approximately the same frequency (+/- 20%)
  - Not too similar to each other (to preserve diversity)
  - Not be similar to known motifs (including the one being shuffled)
- Background level is estimated separately in each region type (e.g. Promoters or 3' UTRs)

#### Branch Length Score ➔ Confidence

1. Use motif-specific shuffled control motifs determine the expected number of instances at each BLS by chance alone or due to non-motif conservation
2. Compute **Confidence Score** as fraction of instances over noise at a given BLS
Motif confidence selects functional instances

Transcription factor motifs

Confidence selects functional regions

Confidence selects in vivo bound sites

Increasing BLS → Increasing confidence

Confidence selects positive strand

Increasing BLS → Increasing confidence

Confidence selects functional regions

Confidence selects functional regions

Confidence selects functional regions

High sensitivity

Kheradpour et al., Genome Research 2007

ChIP vs. conservation: similar power / complementary

Relative functional enrichment

• Together: best
  complementary

• Bound but not conserved:
  reduced enrichment
  Selects functional

• All-ChIP vs. All-cons: similar enr.
  Similar power

• Cons-only vs. ChIP-all: similar
  Additional sites

Kheradpour et al., Genome Research, 2007

Initial regulatory network for an animal genome

ChIP-grade quality

- Similar functional enrichment
- High sens. High spec.

Systems-level
- 81% of Transc. Factors
- 86% of microRNAs
- 8k + 2k targets
- 46k connections

Lessons learned
- Pre- and post- are correlated (hihi/lolo)
- Regulators are heavily targeted, feedback loop

Kheradpour et al., Genome Research, 2007

Network captures co-expression supported edges

Red = co-expressed

Grey = not co-expressed

Named = literature-supported

Bold = literature-supported

46% of edges are supported (P=10⁻³)

Kheradpour et al., Genome Research, 2007

Network captures literature-supported connections

Pre- and primed synaptically control glial transcription of the Tenebrio melanocephala genome

Sushmita Roy

Negative regulation of prominent gene encoding a critical transcriptional repressor of adipocytes

Sushmita Roy

Surprise 3: Abundant feed-forward loops in DV patterning

- Master regulators also bind downstream targets
- Cooperation of master reg. & downstream reg.

Zeitlinger et al., Genes & Development 2007
6. Detecting individual nucleotides under selection

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 site-wise application

Neutral sequence

Decreased rate \(\omega\)

\(\omega_0\) 0 0.5 0.8 1.2 0

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

Unusual patterns \(\pi\)

Individual binding sites are revealed

\(\pi\) log-odds (12mers) \(\omega\) (12mers) \(\omega\) (50mers)

<table>
<thead>
<tr>
<th></th>
<th>12mers</th>
<th>50mers</th>
</tr>
</thead>
<tbody>
<tr>
<td>4 mammals</td>
<td>7.1/1.5/4.6</td>
<td>6.9/1.8/4.1</td>
</tr>
<tr>
<td>29 mammals</td>
<td>5.7/1.1/3.8</td>
<td>5.7/1.8/3.0</td>
</tr>
<tr>
<td>(HMRD) Human Mouse Rat Dog</td>
<td>4.2/0.0/0.0</td>
<td>5.3/1.0/0.3</td>
</tr>
</tbody>
</table>

Increase in power from HMRD to 29 mammals

\(\pi\%\) under constraint constrain means:
That a randomly chosen k-mer in the genome has \(\pi/100\) probability of being under selection.

Comparative genomics I: Evolutionary signatures

- Nucleotide conservation vs. evolutionary signatures
- Protein-coding genes
  - Reading-frame conservation, codon-substitution frequency
  - Likelihood ratio framework: Estimating Q, CQN, scoring
  - Read-through genes, excess constraint regions
- microRNA genes
  - Structural and evolutionary features of microRNAs
  - Combining features: decision trees, random forests
  - Sense/anti-sense miRNAs, mature/star arm cooperation
- Regulatory motifs and motif instances
  - Genome-wide conservation criteria: intergenic, coding, u/d
  - Branch-length score (BLS), control motifs, confidence
  - Inferring regulatory networks and individual binding sites

Manuel Garber, Or Zuk, Xiaohui Xie

Michele Clamp

Over 20 histograms show similar conservation patterns

Or Zuk, Manuel Garber

Manuel Garber, Or Zuk
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