CS273A
The Human Genome Source Code

Gill Lecture 13: Comparative Genomics
Mon, Wed 11:30 AM - 12:50, on Zoom*
Prof: Gill Bejerano
CA: Boyoung (Bo) Yoo
* Track class on Piazza


Announcements

• Hope you’re enjoying the exam

Class Topics

(0) Genome context:
  cells, DNA, central dogma
(1) Genome content / genome function:
  genes, gene regulation, epigenetics, repeats, SARS-CoV-2
(2) Genome sequencing:
  technologies, assembly/analysis, technology dependence
(3) Genome evolution:
  evolution = mutation + selection, main forces of evolution:
  Neutral evolution, Negative selection, Positive selection
(4) Population genomics:
  Human migration, paternity testing, forensics, crypogenomics
(5) Genomics of human disease:
  personal genomics, GxE disease types, deep dive monogenics
(6) Comparative genomics
  Genomics of amazing animal adaptations, ultraconservation

good morning!
Class Topics

1. Genome context:
   cells, DNA, central dogma
2. Genome content / genome function:
   genes, gene regulation, epigenetics, repeats, SARS-CoV-2
3. Genome sequencing:
   technologies, assembly/analysis, technology dependence
4. Genome evolution:
   evolution = mutation + selection, main forces of evolution:
   Neutral evolution, Negative selection, Positive selection
5. Population genomics:
   Human migration, paternity testing, forensics, cryptogenomics
6. Genomics of human disease:
   personal genomics, GxE disease types, deep dive monogenics
7. Comparative genomics:
   Genomics of amazing animal adaptations, ultraconservation

The study of human genetic disease taught us – implicitly –
to appreciate the difference between the health of a
population and the health of an individual in it.

~3% of the human population suffers monogenic diseases.

~1:1,000 baby sea turtles survive to adulthood.

"Nothing in Biology Makes Sense
Except in the Light of Evolution"

Theodosius Dobzhansky
We're not alone in the world…

Even if we stay anthropocentric:

• What's our relationship with other species?
• Whose closest to us?
• How close are they?
• What makes us human?
Comparative Genomics

- Cross species genome comparison

- A powerful way to see the unity of life on earth
  - DNA, RNA, protein
  - Genetic code
  - Gene families
  - Whole genome

The Species Tree

By comparing the genomes of (representatives of) multiple species, we get to see genomic variants that have survived and thrived (fixed) the sieve of selection, in these different species.

Example: Human-Chimp Genomic Differences

Mutations kill functional elements.
Mutations give rise to new functional elements (by duplicating existing ones, or creating new ones)
Selection whittles this constant flow of genomic innovations.

Human and chimp genomes are 96% identical.
(Think about it for a minute)
The basic tenet of comparative genomics – the long version:
Sequence conservation (across large enough evolutionary distances) implies (unknown) genomic function(s) of the conserved region.

purifying selection vs. neutral evolution

Imperfect genome replication bombards the genomes of individuals in the population with all manner of sequence changes. Selection retains/promotes the functional versions/individuals.

Conservation implies function.
Lack of sequence conservation does not imply lack of function.
Lack of seq. conservation does not imply lack of function conservation.
Conservation is sufficient but not necessary to call (unknown) function.
Selection retains/promotes the functional versions/individuals.

Some activity is necessary but not sufficient for function.
E.g., many genes are expressed in many cellular contexts, where knocking them out leads to no observable phenotypes.

The combination of activity assays and sequence conservations leaves us all wanting more.
Best proof of sequence function is via its disruption.
Understanding the reason(s) for sequence conservation is a lot harder.
Conservation as a "magic marker" for function

Comparative Genomics of related species highlights:

functional region!

Conserved elements in the Human Genome

Typical DNA Conservation levels

Conserved elements between human and mouse are on average 85% identical. [Mouse consortium, 2002]
Conserved elements in the Human Genome

Difference: 5% of Human Genome

85% id on average

Ultraconservation

Ultraconserved Elements

481 elements perfectly conserved (100% id) over 200bp or more between human, mouse, and rat.

What exactly is an Ultraconserved Element?

Aha!!
Ultraconservation as a Phenomenon

Few species

More and more species

Ultraconserved Elements: Why?

Hundreds of long genomic regions identical between amniotes. They must have rejected many different changes. But... all functions we understand in our genome are encoded using redundant codes.

Ultras are Functional

Back in 2004 we hypothesized:

481 ultraconserved elements

exonic subset – post transcriptional regulation

"nonsense" subset – transcriptional regulators

Ultras are Under Strong Human Selection

Mutational cold spots? NO. Rare (new) mutations are introduced to the population.
Fierce purifying selection? YES. Very few of these get anywhere near fixation.

[Diagram showing Ultra DAF and NonSyn DAF distributions]

Mutational cold spots? NO. Rare (new) mutations are introduced to the population.
Fierce purifying selection? YES. Very few of these get anywhere near fixation.

Touch an Ultra And You - DIY

[Diagram showing selection elements and Homozygous KIIs]

What can’t we measure in the lab?

To assess the prevalence of phenotypic variation in human population genetic experiments, we measured over 5000 genes with phenotypic data. The database contains 5,000 genes uncorrelated with a human phenotype resulting from human population genetic experiments. 500 genes (12%) gave rise to an unmeasurable phenotype in at least one human experiment. 30 genes (1%) gave rise to an unmeasurable phenotype in any human experiment. Since the human population genetic experiments were conducted in lab, a number of candidate genes were not measured because they were not expressed in the human experiments; and the genetic diversity of the sample population was not considered. The database also contains 1,000 genes with phenotypic variation in at least one human experiment but not measured in the human population genetic experiments. The database is a valuable resource for researchers interested in phenotypic variation and the study of human evolution.

Pr(fixation | N_e, s) = \frac{1 - e^{-s}}{1 - e^{-2N_e s}}

\(N_e\) is population size, a selective disadvantage. Both of which are VERY wrong in the lab.
So it can happen – but does it FIX?

Count Fraction Lost, Binned by %id

Ultras are Fiercely Retained through Evolution
How special are the Ultras?

Ultraconservation identifies a small subset of extremely constrained developmental enhancers and coinhibitory regulatory modules.

Ultraconservation explains how 1% of human DNA maintains its function for ~200 million years across the animal kingdom.

Ultraconservation provides a window on the regulatory instructions of the human genome, specifying the molecular processes that will give rise to each human cell.

Ultraconservation is a phenomenon where few species have conserved regions, and more and more species show an increase in conserved regions.

Hmmm…

We do not see a bump in the curve.

Ultraconservation as a Phenomenon

Question

Is functional density (conservation %id) a good measure of dispensability (deletions)?

Both of these regions are clearly under purifying selection. Which one is more likely to be lost throughout evolution?

[McLean & Bejerano, Genome Res., 2008]
Most all Non coding sequence under purifying selection clearly resists deletions.

- Loss rate nearly constant regardless of conservation %id
- Similar loss rates for slowly-evolving non-exonic sequence and coding exons suggest functionality of non-exonic sequence

If functional density doesn’t correlate with dispensability, what does?

Deep orthology is a better predictor of persistence

Inferring Genomic Histories From Alignments of Genomes
A Gene tree evolves with respect to a Species tree

By “Gene” we mean any piece of DNA.

Terminology

Orthologs: Genes related via speciation (e.g. C, M, H3)
Paralogs: Genes related through duplication (e.g. H1, H2, H3)
Homologs: Genes that share a common origin (e.g. C, M, H1, H2, H3)

Typical Molecular Distances

If they were evolving at a constant rate:
- To which is H1 closer in sequence, H2 or H3?
- To which H is M closest?
- And C?

(Selection may skew distances)
Gene trees and even species trees are figments of our (scientific) imagination.

Species trees and gene trees can be wrong. All we really have are extant observations, and fossils.

**Why Study Homology?**

- **Similar**
  - Sequence
  - Structure
  - Function

Same for RNA genes and gene regulatory elements.

**Gene Families**
I found ultras looking for CNE families..

Group them into paralog families of human functional regions of common origins:
• Annotated members induce function on all.
• Examine core, substitutions in family.
• Test for “guilt by association”. [Bejerano et al., ISMB 2004]

Dotplots for Songs

https://colinmorris.github.io/SongSim/#/

Dotplots
• Dotplots are a simple way of seeing alignments
  • We really like to see good visual demonstrations, not just tables of numbers
  • It’s a grid: put one sequence along the top and the other down the side, and put a dot wherever they match.
  • You see the alignment as a diagonal
• Note that DNA dotplots are messier because the alphabet has only 4 letters
  • Smoothing by windows helps.
Sequence Alignment

AGGCTATCCCTCAGCTCCAGGGCGATGCCC
TAGCTATACGACCCGCGGATTTGCCGAC

~AGGCTATCCCTCAGCTCCAGGGCGATGCCC
~TAGCTATACGACCCGCGGATTTGCCGAC

Definition
Given two strings \( x = x_1x_2\ldots x_n \) and \( y = y_1y_2\ldots y_m \), an alignment is an assignment of gaps to positions \( 0, \ldots, n \) in \( x \), and \( 0, \ldots, m \) in \( y \), so as to line up each letter in one sequence with either a letter, or a gap in the other sequence.

Scoring Function

- Sequence edits: AGGCCTC
- Mutations AGGACTC
- Insertions AGGGCTTC
- Deletions AGG CTC

Score \( F \) = (\# matches) \cdot m - (\# mismatches) \cdot s - (\# gaps) \cdot d

Scoring Function

<table>
<thead>
<tr>
<th>Operation</th>
<th>Cost</th>
</tr>
</thead>
<tbody>
<tr>
<td>Match</td>
<td>+m</td>
</tr>
<tr>
<td>Mismatch</td>
<td>-s</td>
</tr>
<tr>
<td>Gap</td>
<td>-d</td>
</tr>
</tbody>
</table>

Alternative definition:

Cost of edit operations needs to be biologically inspired (e.g., DEL length).

Solve via Dynamic Programming

Are two sequences homologous?

AGGCTATCCCTCAGCTCCAGGGCGATGCCC
TAGCTATACGACCCGCGGATTTGCCGAC

~AGGCTATCCCTCAGCTCCAGGGCGATGCCC
~TAGCTATACGACCCGCGGATTTGCCGAC

Given an (optimal) alignment between two genome regions, you can ask what is the probability that they are (not) related by homology?

Note that (when known) the answer is a function of the molecular distance between the two (e.g., between two species).
Sequence Alignment

AGGCTATCACCTGGACTCCAGGCCGATGCC
TAGCTATCACGCCCGTGGATTTGCGGAC

AGGCTATCACCTGGACTCCAGGCCGATGCC
TAGCTATCACGCCCGTGGATTTGCGGAC

Similarity is often measured using "%id", or percent identity

%id = number of matching bases / number of alignment columns

Where

Every alignment column is a match / mismatch / indel base

Where indel = insertion or deletion (requires an outgroup to resolve)

Note the pattern of sequence conservation / divergence


"Raw" (B)lastz track (no longer displayed)
Chaining co-linear alignment blocks

Objective: find local alignment blocks, that are likely homologous (share common origin)

Chaining strings together co-linear blocks in the target genome to which we are comparing.
Double lines when there is unalignable sequence in the other species. Single lines when there isn’t.

Conservation Track Documentation

Chaining Alignments

Chaining highlights homologous regions between genomes, bridging the gap between syntetic blocks and base-by-base alignments.

Local alignments tend to break at transposon insertions, inversions, duplications, etc.

Global alignments tend to force non-homologous bases to align.

Chaining is a rigorous way of joining together local alignments into larger structures.
Chaining Algorithm

Input - blocks of gapless alignments from (b)lastz
Dynamic program based on the recurrence relationship:
\[ \text{score}(B_i) = \max(\text{score}(B_j) + \text{match}(B_i) - \text{gap}(B_i, B_j)) \]

Uses Miller's KD-tree algorithm to minimize which parts of dynamic programming graph to traverse. Timing is \(O(N \log N)\), where \(N\) is number of blocks (which is in hundreds of thousands).


“Raw” (B)lastz track (no longer displayed)

Chains join together related local alignments
Before and After Chaining

<table>
<thead>
<tr>
<th></th>
<th>Blastz</th>
<th>AxtChain</th>
</tr>
</thead>
<tbody>
<tr>
<td>Longest</td>
<td>63 k</td>
<td>115 M</td>
</tr>
<tr>
<td>Average</td>
<td>608</td>
<td>23 k</td>
</tr>
<tr>
<td>Count</td>
<td>8.6 M</td>
<td>147 k</td>
</tr>
</tbody>
</table>

Nets/chains can reveal retrogens (and when they jumped in!)

Interspersed vs. Simple Repeats

From an evolutionary point of view transposons and simple repeats are very different.

Different instances of the same transposon share common ancestry (but not necessarily a direct common progenitor). Different instances of the same simple repeat most often do not.
Note: repeats are an alignment nuisance

If, for example, human and mouse have each 10,000 copies of the same repeat. We will obtain and need to output 10^8 alignments of all these copies to each other.

Note that for the sake of this comparison interspersed repeats and simple repeats are equal nuisances. However, note that simple repeats, but not interspersed repeats, violate the assumption that similar sequences are homologous.

Solution:
1. Discover all repetitive sequences in each genome.
2. Mask them when doing genome to genome comparison.
3. Chain your alignments.
4. Add back to the alignments only repeat matches that lie within pre-computed chains.
   This re-introduces back into the chains (mostly) orthologous copies.
   (which is valuable!)

Nets attempt to computationally capture orthologs

Nets find best match mouse match for each human region.
Highest scoring chains are used first.
Lower scoring chains fill in gaps within chains inducing a natural hierarchy.

Netting Alignments

Commonly multiple mouse alignments can be found for a particular human region, e.g. including for most coding regions.
Net finds best match mouse match for each human region.
Highest scoring chains are used first.
Lower scoring chains fill in gaps within chains inducing a natural hierarchy.
Before and After Netting

<table>
<thead>
<tr>
<th></th>
<th>Blastz</th>
<th>AxtChain</th>
<th>Net Top</th>
</tr>
</thead>
<tbody>
<tr>
<td>Longest</td>
<td>63 k</td>
<td>115 M</td>
<td>115 M</td>
</tr>
<tr>
<td>Average</td>
<td>608</td>
<td>23 k</td>
<td>147 k</td>
</tr>
<tr>
<td>Count</td>
<td>8.6 M</td>
<td>147 k</td>
<td>20 k</td>
</tr>
</tbody>
</table>

Convert / LiftOver

“LiftOver chains” are actually chains extracted from nets, or chains filtered by the netting process.

What nets can’t show, but chains will
Same Region...

same in all the other fish

Self Chain reveals (some) paralogs

(self net is meaningless)