Ultraconservation & Function in the Human Genome

Gill Bejerano
School of Engineering, UC Santa Cruz

Gill Bejerano, UCSC  

w: www.soe.ucsc.edu/~jill
Functional Genomics

Objective: Understanding the function and evolution of every base in the human genome.
Talk Outline

- The Annotated Human Genome
- The Comparative Genomics Paradigm
- Families and Functions of Conserved ncDNA
- Ultraconservation in Mammals, Vertebrates, and Beyond
- Summary
Dude, Where’s My Complexity?

Organism complexity does not correlate with neither
- Genome size
- Number of genes

E.g.,

<table>
<thead>
<tr>
<th>Genome</th>
<th>Genes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Human</td>
<td>3Gb</td>
</tr>
<tr>
<td>C.elegans</td>
<td>100Mb</td>
</tr>
</tbody>
</table>

What should explain it then?
- Alternative splicing
- Gene expression control
- Chromatin remodeling
- RNA regulatory networks
- Transposable elements

Gill Bejerano, UCSC
The Annotated Human Genome

- ~2.9Gb
- coding exons 1.5%
- known function 0.5%
- repeats 50%
- other 48%
- ~2.9Gb
- functional

Prediction tools
- good: protein coding genes, repeats*
- improving: alt-splicing, TF binding sites, ncRNA, RNA editing
- infancy: enhancers, silencers, insulators, S/MARs
Comparative Genomics

Functional DNA often evolves slower than neutral DNA.

To detect functional elements:
align genomes of related species,
and find regions of high conservation.

E.g.,

http://genome.ucsc.edu
≥5% of H.G. Appears Under Selection!

compare to mouse -
40% DNA alignable
95% coding genes shared

H.G.
2.9Gb

coding exons 1.5%
known function 0.5%
other 48%
repeats 50%

neutral human-mouse DNA
all human-mouse DNA

Difference: 5% of Human

better get to work...

[Mouse Consortium 2002]
Functional Annotation by Paralogy

Removed from H.G. top 5% all annotated regions, and then some:

GeRUF = Genomic Region of Unknown Function

Group them into families of human paralogs.
- Annotated members induce function on all.
- Examine functional core of family.
- Test for “guilt by association”

700,000
3.5% H.G.

(TABLE 1). Remarkably, these sequences are not repetitive and do not share features that are identifiable by primary sequence comparisons. BLAST searches of all CNGs against the whole human genome fail to identify significant matches except in the case of old paralogues that have resulted from segmental duplications. (This is true even if conserved sequences are defined with different criteria,

[Dermitzakis et al., Nat. Rev. Gen. 2005]
Functional Annotation by Paralogy

Removed from H.G. top 5% all annotated regions, and then some:

GeRUF = Genomic Region of Unknown Function

Group them into families of human paralogs.
• Annotated members induce function on all.
• Examine functional core of family.
• Test for “guilt by association”

Bad News:
96% of 700,000 GeRUFs have no paralogs.

Good News:
We still find 12,027 families!

[Bejerano et al., ISMB 2004]
GeRUF Clustering

1) Define the set of intervals of interest.
2) Define a similarity measure between any pair of intervals.
3) Build a weighted similarity graph, where Vertices = GeRUFs, and Edge weights = pairwise similarity.
4) Find biologically motivated dense subgraphs.
5) Analyze and annotate clusters.
   (Refine and reiterate…)

Human.chr1  GCACC...AGCGC
Human.chrX  ATAGA...GCACA
Using Sean Eddy’s track of known human RNA genes we find:

- 47 clusters containing only known RNA genes
- 30 clusters containing some known RNA genes
- 50 clusters with (mfold) conserved secondary structure in many/all members, but no annotation, e.g.:
Jumping Alt Spliced Exon

Element overlaps exon in:
- PCBP2: poly C binding protein 2
- SMARCA4: member, SWI/SNF family
- ATF2: activating T.F. 2
- Other proteins of diverse functions

It also occurs in an intron or upstream of other unrelated genes.
Iroquois (IRX) Gene Clusters

Early body patterning

Ancient duplications – Drosophila has one such clusters
Chicken has both

Hs.chr16

IRX-3 650 kb IRX-5 390 kb IRX-6

Hs.chr5

IRX-1 840 kb IRX-2 890 kb IRX-4

35 GeRUFs

IRX genes bracket 4 gene deserts*

45 GeRUFs

Gill Bejerano, UCSC
IRX Co-regulation / Insulation?

- Early body patterning

Hs.chr16

IRX-3

IRX-5

IRX-6

70-80%id

Hs.chr5

IRX-1

IRX-2

IRX-4

Gill Bejerano, UCSC
No known function requires *this* much conservation!

contamination?!
Contamination

An issue in initial genome drafts (sample switching, vector, ...)

We employed ample defenses:
• Human and Mouse are high quality drafts.
• The addition of the Rat genome.
• Syntenic nets.
**Ultra Conserved (UC) Elements**

*Any* contiguous block of human-mouse-rat alignment that is identical in all three species, syntenic and ≥200 bp. * annotated or not

(p=10^-22 of finding one such element in slowest rate 2.9G neutral DNA)

Turns out there are **481(!) such blocks of sizes 200-779bp**

(total of 126Kb) in all chromosomes but 21, Y.

*68 (61%) associated with alt-spliced exons*
By joining two ultras into a cluster when separated <675Kb we obtained 89 clusters (each named after prominent gene/s)

Non exonic elements tend to congregate in clusters.

Exonic elements are distributed more randomly (tend to overlap an alt-spliced exon).
Genomic Distribution

- exonic
- non
- possibly
Functional Annotation of Related Genes

Exonic – RNA processing @ transcription regulation
Non Exonic – regulation of transcription at DNA level
Subs. rate 20-fold reduced for >300Mya

Independent species comparison

<table>
<thead>
<tr>
<th>species (%Hs covered)</th>
<th>matching elements</th>
<th>identical bases</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chicken (4%)</td>
<td>467 (97%)</td>
<td>95.7%</td>
</tr>
<tr>
<td>Fugu (1.8%)</td>
<td>324 (67%)</td>
<td>76.8%</td>
</tr>
</tbody>
</table>

(29 perfectly)

Single base differences (excl. 2x20bp)

<table>
<thead>
<tr>
<th></th>
<th>observed</th>
<th>expected</th>
<th>under rep.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Validated SNPs</td>
<td>6 / 106,767</td>
<td>119</td>
<td>20x (p=10^{-42})</td>
</tr>
<tr>
<td>Hi Qual chimp diffs</td>
<td>38 / 106,767</td>
<td>716</td>
<td>19x (p=10^{-200})</td>
</tr>
</tbody>
</table>
Non Exonic Enhancers

The non exonic ultras are often found in “gene deserts” (140 / 256 > 10Kb from a known gene; 88 > 100Kb away).

The genes flanking these ultras are GO enriched for development (p = 10^{-6}), particularly early developmental tasks (p = 2-7 \times 10^{-5}) suggesting distal enhancer roles.

Indeed, **uc.351** is contained in a proven enhancer of DACH, located 225Kb upstream of it [Nobrega et al., *Science* 2003].
Zoom to uc.351, 225Kb upstream of DACH
The most conserved elements in the genome

If one concatenates ultras that are 1-2bp away, all 4 longest ultras (1044, 779, 731, 711bp) lie at 3’ end of POLA, near ARX, on chr X. The longest has 8 subs. and no indels (99.3%id) in chicken.

ARX is a homeobox gene involved in CNS development; defects in the gene are linked to epilepsy, mental retardation, autism and cerebral malformations.
Exonic uc’s correlate with Alt Splicing

68 / 111 exonic ultras overlap an exon that shows clear evidence of being alternatively spliced.

Of the 59 GO annotated genes containing these elements:
- 24 are RNA binding ($p = 8.1 \times 10^{-18}$), including HNRPU, HNRPD1, HNRPH1, HNRPK, HNRPM.
- 16 contain the RNA recognition motif ($p = 9.1 \times 10^{-19}$), including SFRS1, SFRS3, SFRS6, SFRS7, SFRS10, SFRS11.

These ultras often overlap a short exon that is retained only in some tissues.

Such is the explicitly studied, \textbf{uc.33} in PTBP2 (length 312bp) overlaps a 34bp exon included in the mature transcript only in the brain [Rahman et al., \textit{Genomics} 2004]
A Vertebrate Innovation?

Only 24 ultras can be partially traced back through direct sequence search to *Ciona, C. Elegans* or *Drosophila*. All overlap coding exons from known genes (17 of which show clear evidence of alt-splicing inc. EIF2C1, DDX, BCL11A, EVI1, ZFR, CLK4, HNRPH1, GRIA3).

No intronic element in human was found to be coding in another species, although in some cases EST evidence indicates intron retention, presumably not as CDS.

Interestingly, **ribosomal DNA** (not part of the draft genomes) also harbors 6 ultra conserved elements in 18S, 28S.
Define Ultra…

  ≥100bp, 100%id – 5000 human-mouse-rat elements.

* 417 Human-Chicken ultras [≥200bp identical, syntenic]
  Only 114/417 (27%) overlap the mammalian ultras!
  However, all chicken ultras are in mouse/rat.
  Very few SNPs. No homology beyond vertebrates.
  Non-exonics enriched for TF, DNA binding, Homeobox.

<table>
<thead>
<tr>
<th>Ultras</th>
<th>w/Chicken</th>
<th>w/Rodents</th>
</tr>
</thead>
<tbody>
<tr>
<td>exonic</td>
<td>33/417</td>
<td>111/481</td>
</tr>
<tr>
<td></td>
<td>(8%)</td>
<td>(23%)</td>
</tr>
<tr>
<td>non-exonic</td>
<td>284/417</td>
<td>256/481</td>
</tr>
<tr>
<td></td>
<td>(68%)</td>
<td>(53%)</td>
</tr>
<tr>
<td>non-exonic</td>
<td>137/417</td>
<td>100/481</td>
</tr>
<tr>
<td>intronic</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(33%)</td>
<td>(21%)</td>
</tr>
</tbody>
</table>

[Chicken Consortium 2004]
Key Mysteries

• What molecular mechanisms lead to ultra conservation?
• Why the strong association with DNA and RNA binding genes? Are ultras part of delicate self-regulating gene networks that are critical in development?
• Did transposons originally “seed” the ultras, and if so, when?
• How do ultras evolve: all by purifying selection, or is there some reduced mutation or hyper-repair?

Hypotheses

• Strong negative selection (heterozygous disadvantage)
• Multiple functions (a-la ribosomal RNA)
• Hypo-mutation / hyper-repair (neutral DNA / functional core)
• Gene conversion / Convection related (rRNA, yeast Mat)

is the phenomenon restricted to vertebrates?
Fly Ultras

Flies have (insect-specific, shorter) ultras.
Dm-Dp: $841 \geq 100\text{bp}$ ($11 \geq 200\text{bp}$)
Dm-Dp-Am: $87 \geq 50\text{bp}$.
Non-exonics enriched for “trans-dev” (as in vertebrates).
Longest ultras are exonics, enriched in ion channels.

Longest Dm-Dp-Ag ultra is exonic (exon-intron jct.) in the \textit{hth} (homothorax) TF.
Tested in the lab – mediates intron retention.
Folds into hairpin structure.
Not found in Bee.

[w/Mattick group]
More Species, More Elements

Two state Phylo-HMM

\[ \gamma = \frac{\nu}{\mu + \nu} \]
\[ \omega = \frac{1}{\mu} \]

\[ X = \text{TCGCGACATATACGA...} \]
\[ \text{TTGCGCATGTTGCTCA...} \]
\[ \text{AGCAGACGTCCGCAA...} \]

[w/Adam Siepel]

Gill Bejerano, UCSC
Coverage / Composition

Vertebrate (H. sapiens)

Fly (D. melanogaster)

Worm (C. elegans)

Yeast (S. cerevisiae)

[w/Adam Siepel]
Different Notions of “Most” Conserved

Prefers longer, highly conserved to shorter perfectly conserved.

Flushes out things like extreme conservation of 3’ UTRs, enriched in RNA binding proteins, in both vertebrates and flies.

**Protocol** for grabbing and analyzing conserved elements. (focuses on screening for putative enhancers)  

[w/Adam Siepel]
Functional Assignment

Phylo-SCFG
RNA structure prediction
[w/Jakob Pedersen]

Understanding Enhancers
[w/Marcelo Nobrega]

ultraconservation - too much of a good thing...
Summary

• Many many conserved elements out there.
• Think families.
  Different kinds, different metric, different clustering.
• Ultraconservation comes in more than one flavour
  – Non-exonics: enhancers
  – Exon-intron junctions: involved in splicing
  – What else? ncRNAs? S/MARs?!
• Ultras are not restricted to vertebrates.
  Seem to correlate with more complex metazoans(?)
Kudos

**UCSC**
David Haussler
Jim Kent, Sofie Salama
Adam Siepel, Jakob Pedersen
Katherine Pollard, Craig Lowe
Hiram Clawson, Angie Hinrichs
Genomics/Browser Group

Experimental Collaborators*
Eddy Rubin (LBL)
  N. Ahituv, M. Nobrega, D. Boffelli
*Rick Myers (Stanford)*
*Richard Harland (Berkeley)*
Bruce Roe (U. Oklahoma)

**McGill U.**
Mathieu Blanchette

**U. Queensland**
John Mattick’s group

**Penn State U.**
Webb Miller

Genome Sequencing Consortia

*Additional collaborations welcomed!*