This is “the Century of Biology”

EDITORIAL

Unification in the Century of Biology

Scientific progress is based ultimately on unification rather than fragmentation of knowledge. At the threshold of what is widely regarded as the century of biology, the life sciences are undergoing a profound transformation. They have long existed as a collection of narrow, even parochial, disciplines with well-defined territories. Now they are undergoing consolidation, forming two major domains: one extending from the molecule to the organism, the other bringing together population biology, biodiversity studies, and ecology. Kept separate, these domains, no matter how fruitful, cannot hope to deliver on the full
We can now cast Biology in “our” terms

strings

circuits

time series

http://bejerano.stanford.edu
The Meaning of Life (abridged)
DNA: Functional and Non-Functional

DNA = linear molecule that carries genetic instructions for making living organisms ~ long string over a small alphabet
Alphabet of four \{A,C,G,T\} Strings of length $10^4$-$10^{11}$

...ACGTACGACTGACTAGCATCGACTACGACTAGC...

“junk” DNA

genetic instructions:
how to...
when to...
where to...

“junk” DNA
One Cell, One Genome, One Replication

Every cell holds a single copy of all its DNA = its genome. The genome is replicated every cell division. The human body is made of $\sim10^{14}$ cells. All originate from a single cell through cell division.

http://bejerano.stanford.edu
“Nothing in Biology Makes Sense Except in the Light of Evolution”

Theodosius Dobzhansky
DNA Replication is Imperfect

Small Scale: single letters are substituted, erased, added

Small change: ...ACGTACGACT... becomes ...ACGTACGACTAGCATCGACTACGA...

Chicken egg: functional

“anything goes”: many changes are not tolerated

Thus, sequence conservation $\Rightarrow$ function!
Conservation implies Function

Comparative Genomics of Distantly related species:

functional region!

http://bejerano.stanford.edu

(but which function/s?...)

human

mammalian ancestor

mouse

http://bejerano.stanford.edu
The Human Genome is Full of Mysteries

all human-mouse DNA
human-mouse junk DNA

Difference: 5% of Human

Human Genome: 3*10^9 letters

1.5% known function

>50% junk

why bother?

3x more functional DNA than known!

But what do these 10^7 substrings do?..

[Mouse Consortium 2002]
[Science 2004 Breakthrough of the Year, 5th runner up]
Genes, Proteins and Gene Control

DNA

proximal: in $10^3$ letters

Translation

gene (how to)

control region

(when & where)

proteins

3kb

genome.ucsc.edu

http://bejerano.stanford.edu
Ultraconserved Elements

<table>
<thead>
<tr>
<th>Base Position</th>
<th>71566500</th>
<th>71567000</th>
<th>71567500</th>
<th>71568000</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chromosome Band</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Conservation</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>chimp</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>dog</td>
<td></td>
<td></td>
<td></td>
<td></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>chicken</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>fugu</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>zebrafish</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

RefSeq Genes
Hu/Chimp/Mouse/Rat/Dog/Chick/Fugu/Zfish Multiz Alignments & Conservation

Ultraconserved Elements (200 bp 100% ID in rat/mouse/human)
uc.381

Human/Chimp/Mouse/Rat/Chicken/Fugu/Zebrafish Multiz Alignments & PhyloHMM Cons

HOXA4 exon

[Bejerano et al., Science 2004]
Why is Perfect Conservation So Surprising?

If a substring is identical between enough distant species, it must have rejected many different changes over time. But... **all** functions we understand in our genome are encoded using **redundant codes**.

E.g. Protein Coding Genes:

- DNA – $10^8$ letters over alphabet of 4.
- Protein – $10^2$ letters over alphabet of 20.

Coding: 3 DNA letters $\rightarrow$ 1 Protein letter.
Genes, Proteins and Gene Control *Revisited*

Gene (how to) control region (when & where)

DNA

- distal: in $10^6$ letters
- proximal: in $10^3$ letters

Translation

DNA binding proteins

protein
Vertebrate Gene Regulation

- gene (how to) control region (when & where)

- ~10^6 letters!!!

- DNA

- ~10^3 letters

- protein

- crucial regulation
- many thousands
- previously invisible

http://bejerano.stanford.edu
Ultraconserved Elements

481 regions *perfectly* conserved over 200 DNA bases or more, between human, mouse and rat (P<10\(^{-22}\) in "junk")

- Evolve 20-fold slower than human average.
- Most do *not* overlap protein coding DNA.
- Those that do not code cluster spatially, near genes encoding DNA binding proteins. Dozens validated since as controlling genes.
- Those that do code, are found in genes coding for a specific type of protein.
- The tip of a continuum of very slowly evolving elements.
- The ultras cannot be found beyond vertebrates.

[Bejerano et al., *Science* 2004
Origins of Ultraconserved Elements?

http://bejerano.stanford.edu
Origins of Ultraconserved Element

http://bejerano.stanford.edu
Coelacanth Homologs to uc.338 Closer than Human Ones

[Bejerano et al., Nature 2006]

http://bejerano.stanford.edu
Coelacanth “the Living Fossil” Fish

Fossil Record: Appeared >360Mya, Peaked 240Mya, Disappeared 80Mya
Rediscovered (by science) in 1938. Possible Explanation: Habitat Switch.

http://bejerano.stanford.edu
Repeats / mobile Elements ("selfish DNA")

Human Genome: 3\times10^9 letters

1.5% known function

>50% junk

http://bejerano.stanford.edu
Upto 80% id between Coelacanth repeat and human instances, inc uc.338.
Cis-reg & Ultra elements from Mobile Elements

Co-option event, probably due to favorable genomic context

All other copies are destined to decay over time at a neutral rate

http://bejerano.stanford.edu

[Yass is a small town in New South Wales, Australia.]

[Bejerano et al., Nature 2006]
Exapted Into Which Cellular Roles?

No evidence for Transcription (Tx) as small RNAs, no orientation preference in introns, not in antisense Tx.

<table>
<thead>
<tr>
<th>Organism</th>
<th>5' UTR</th>
<th>3' UTR</th>
<th>Exonic Alt-Spliced</th>
<th>Total</th>
<th>Intronic</th>
<th>Intergenic</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>Homo sapiens</em></td>
<td>1</td>
<td>0</td>
<td>12</td>
<td>13</td>
<td>68</td>
<td>163</td>
<td>245</td>
</tr>
<tr>
<td><em>Pan troglodytes</em></td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>210</td>
</tr>
<tr>
<td><em>Macaca mulatta</em></td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>229</td>
</tr>
<tr>
<td><em>Canis familiaris</em></td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>235</td>
</tr>
<tr>
<td><em>Bos taurus</em></td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>169</td>
</tr>
<tr>
<td><em>Mus musculus</em></td>
<td>0</td>
<td>1</td>
<td>7</td>
<td>8</td>
<td>25</td>
<td>57</td>
<td>91</td>
</tr>
<tr>
<td><em>Rattus norvegicus</em></td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>87</td>
</tr>
<tr>
<td><em>Monodelphis domestica</em></td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>394</td>
</tr>
<tr>
<td><em>Gallus gallus</em></td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>244</td>
<td>451</td>
<td>699</td>
</tr>
<tr>
<td><em>Xenopus tropicalis</em></td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>10</td>
<td>14</td>
<td>26</td>
</tr>
</tbody>
</table>

Human instances cluster together, found <1Mb from 35 TFs (P<3*10^{-6}).

http://bejerano.stanford.edu
Transient Transgenics

Eddy Rubin’s Lab, LBNL

Construct is injected into 1 cell embryos
Taken out at embryonic day 10.5-14.5
Assayed for reporter gene activity
ISL1 is a neuro-developmental gene, also expressed in testis. Three previously known enhancers are conserved in all vertebrates.
Mouse Isl1 *in situ* (B) vs. LacZ driven by LF SINE region (C)

- Matched staining in dorsal apical ectodermal ridge (part of limb bud)
- Matched staining in genital eminence

Nadav Ahituv, Eddy Rubin
Corresponding expression patterns in:
(a, b) the developing thalamus (Th) and basal plate (BP) in the brain.
(c, d) the trigeminal (V) ganglion and facio-acoustic (VII/VIII) ganglia in the head region.
(e, f) the dorsal root ganglion (DRG), and the lateral region of the ventral horn (VH) of the spinal cord in thoracic sections.

Bryan King, Sofie Salama, Nadav Ahituv, Eddy Rubin
DNA Replication is Imperfect (contd)

Medium Scale: substrings are duplicated, deleted, inverted
Large Scale: whole DNA strings are duplicated, deleted

junk

...ACGTACGACTGACTAGCATCGACTACGA...

functional

...ACGTACGACTGACTAGCATCGACTACGA...

substring duplication

functional

functional

...ACGTACGACTGACTAGCATCGACTACGA...TCTGACTAGCATCGACTACGA...

functional divergence

functional

functional’

...ACGTACGACTGACTAGCATCGACTACGA...TCTGACTAGCATCGACTACGA...

functional”

...ACGTACGACTGACTAGCATCGACTACGA...TCTGACTAGCATCGACTACGA...

So...More Genes...More Complexity!!...Right?
Genes & Complexity

Gene numbers do not correlate with organism complexity. Many gene families are surprisingly old.

pre-genomic era: “100,000 genes to the human genome”
The Evolution of Morphological Diversity

Gene numbers do **not** correlate with organism complexity. Many gene families are surprisingly old.

“Regulatory sequence evolution must be the major contribution to the evolution of form.” [Sean Carroll, *PLoS Bio* 2005]

---

Hold on... *junk* DNA can contribute these elements

http://bejerano.stanford.edu
From junk DNA to recruitment into pathway?

1) A portion of the genome containing a new remnant realization:

2) "Diffusion" of sequences throughout the genome by subsequent chromosomal rearrangements:

3) Among some local arrangements which might thus arise could be these:

4) In this way new regulatory pathways could arise, for example:

[Davidson & Erwin, 2006]

[Britten & Davidson, 1971]
Same Junk, Different Functional Elements

- Protein coding
- Repeat
- Gene regulating

http://bejerano.stanford.edu
Additional Mysteries Abound
Genome in Flux

Human Genome

Copied out to make protein coding genes

Copied out to make ... ???

Origins & Evolution

Functions & Encoding

Contribution to Human Disease

http://bejerano.stanford.edu
Bejerano Lab: Research Interests


Origins & Evolution

Functions & Encoding

Contribution to Human Disease

[Ernst Haeckel, 1866]
Many thousands of human conserved elements congregate en-masse near developmental genes.

Origins & Evolution

Functions & Encoding

Contribution to Human Disease

Break regulatory code

• syntax
• grammar
• meaning
Bejerano Lab: Research Interests

Many thousands of human conserved elements congregate en-masse near developmental genes.

Origins & Evolution

Functions & Encoding

Contribution to Human Disease

Understand our evolution
- Reconstruct ancient genomes
- Track regulatory regions histories

In/vertebrate Divide

http://bejerano.stanford.edu
Many thousands of human conserved elements congregate en-masse near developmental genes.

Origins & Evolution
Functions & Encoding
Contribution to Human Disease

Make a difference
• “bench to bedside”
Many thousands of human conserved elements congregate en-masse near developmental genes.

**Origins & Evolution**

**Functions & Encoding**

**Contribution to Human Disease**

**Discovery tools**

- large databases
- heterogeneous, noisy data
- statistical correlations
- human interfaces

thousands and thousands of page requests served *daily*

exponential growth of public data

http://bejerano.stanford.edu
Summary

We are only beginning to understand the complexity unearthed by observing whole genomes.

Technology (genome sequencing, gene chips, etc) isflooding us with different form of whole genome measurements – extremely valuable, if challenging.

Some of the challenges discussed today:

• Explain Ultraconservation in particular, and the myriad of unexplained constrained elements in our genome.
• Understand the evolution of morphological diversity (how much has repeats contributed to it quantitatively and qualitatively)
• Understand why so much of our genome is transcribed.
Kudos

UC Santa Cruz
David Haussler
Sofie Salama, Jim Kent, Craig Lowe,
Bryan King, Adam Siepel, Jakob Pedersen
Katie Pollard, Courtney Onodera
Rachel Harte, Genomics/Browser Group

Lawrence Berkeley Labs
Eddy Rubin
Nadav Ahituv

McGill U.
Mathieu Blanchette

Penn State U.
Webb Miller’s group

U. Queensland
John Mattick’s group

Genome Sequencing Consortia
All GenBank contributors

Gill Bejerano
bejerano@stanford.edu

http://bejerano.stanford.edu