Dark Matter of the Human Genome

Gill Bejerano
Dept. of Developmental Biology
Dept. of Computer Science
Stanford University

http://bejerano.stanford.edu
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
We can Harness It

Bioengineering
Synthetic Biology

Embryonic Development

one cell

organism

Enter DNA ...

http://bejerano.stanford.edu
DNA = linear molecule that carries instructions for making living organisms ~ long string(s) over a small alphabet
Alphabet of four \{A,C,G,T\} Strings of length $10^4$-$10^{11}$
One Cell, One Genome, One Replication

Every cell holds a copy of all its DNA = its genome.
The genome is replicated every cell division.
The human body is made of \( \sim 10^{14} \) cells.
All originate from a single cell through repeated cell divisions.

http://bejerano.stanford.edu
Genes = How to make Proteins

“the workhorses of every living cell”
DNA Replication is Imperfect

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

junk \hspace{1 cm} functional

...ACGTACGACTGACTAGCATCGACTACGA...

substrings duplication

functional

...ACGTACGACTGACTAGCATCGACTACGA

functional divergence

functional'

...ACGTACGACTGACTAGCATCGACTACGA

functional'' functional'

...ACGTACGACTGACTAGCATCGACTACGA

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

http://bejerano.stanford.edu
1. Gene number does **not** correlate with Complexity

Gene families are important. Many are surprisingly old. But -

pre-genomic era: “100,000 genes to the human genome”

http://bejerano.stanford.edu
DNA Replication is Imperfect (contd)

Small Scale: single letters are substituted, erased, added

junk

TT

functional

CAT

“anything goes”

many changes are not tolerated

thus, sequence conservation over generations implies function!
Sequence Conservation implies Function

Comparative Genomics of Distantly related species:

functional region!

human → mammalian ancestor → mouse

...CTT[TGCGA−TGAGTAGCATCTACTA]TTT...

...ACG[TGGGACTGACTA−CATCGACTA]CGA...

(but which function/s?...)

http://bejerano.stanford.edu
2. Human Genome full of Conserved Non-Coding Elements

- Human Genome: 3*10^9 letters
  - 1.5% known function
  - >50% junk

- >5% human genome functional

- 3x more functional DNA than known!
- ~10^6 substrings do not code for protein

What do they do then?

[Science 2004 Breakthrough of the Year, 5th runner up]
Gene regulation = when/where to make protein

recognition site ~10 letters/protein

gene (how to)

control region (when & where)

effective region ~10^3 letters

DNA

Translation

protein
Vertebrate Gene Regulation

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

effective region \( \sim 10^6 \) letters!!!

DNA

\((\sim 10^3 \text{ letters})\)

Translation

protein
3. Most Non-Coding Elements are likely cis-regulatory

“IRX1 is a member of the Iroquois homeobox gene family. Members of this family appear to play multiple roles during pattern formation of vertebrate embryos.”
4. Regulatory regions *drive* morphological diversity

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

The Writing on the Wall…

- **gene deserts**
- **regulatory jungles**

http://bejerano.stanford.edu
A Computational Question

Classical Biological approach: experiment to understand these regions
Computational approach: how many similar regions or better are there?

human

related elements
(75% id over 200bp)

related genes

mouse

rat

same element
96% id over 200bp

same element
95% id over 200bp

http://bejerano.stanford.edu
Hundreds of long substrings identical between human-birds ➔ they must have rejected many different changes.

But... all functions we understand in our genome are encoded using redundant codes.

http://bejerano.stanford.edu

[Bejerano et al., Science 2004] 19
Ultraconserved Elements

*Hundreds* of long substrings *identical* between human-birds ⇒ they must have *rejected many different changes.* 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 $→$ 1 Protein letter.

http://bejerano.stanford.edu

[Bejerano et al., *Science* 2004]
Computational Hypotheses

Based on public domain genome wide data:

ultraconserved elements

one subset codes protein

generate testable hypotheses for function from existing knowledge (2004)

[Pennacchio et al., Nature, 2006]

http://bejerano.stanford.edu
Validating Regulatory Elements

Conserved Element → Minimal Promoter → Reporter Gene

where is the reporter gene expressed?

where is the wild type gene expressed?

transgenic → wild type
Origins of Ultraconserved Elements?
Genomic Distribution of Ultraconserved Elements

• exonic
• non
• possibly

Origins?
Uniquely Abundant in Coelacanth

Upto 80% id between Coelacanth instances and some human instances, inc uc.338.

<table>
<thead>
<tr>
<th>Species</th>
<th>UCSC Assembly</th>
<th>LF-SINE Detected</th>
<th>Species</th>
<th>UCSC Assembly</th>
<th>LF-SINE Detected</th>
</tr>
</thead>
<tbody>
<tr>
<td>Homo sapiens</td>
<td>hg17</td>
<td>Yes</td>
<td>Danio rerio</td>
<td>danRer2</td>
<td>No</td>
</tr>
<tr>
<td>Pan troglodytes</td>
<td>panTro1</td>
<td>Yes</td>
<td>Tetraodon nigroviridis</td>
<td>tetNig1</td>
<td>No</td>
</tr>
<tr>
<td>Macaca mulatta</td>
<td>rheMac1</td>
<td>Yes</td>
<td>Takifugu rubripes</td>
<td>fr1</td>
<td>No</td>
</tr>
<tr>
<td>Mus musculus</td>
<td>mm6</td>
<td>Yes</td>
<td>Ciona intestinalis</td>
<td>cil</td>
<td>No</td>
</tr>
<tr>
<td>Rattus norvegicus</td>
<td>rn3</td>
<td>Yes</td>
<td>Strongylocentrotus purpuratus</td>
<td>strPur1</td>
<td>No</td>
</tr>
<tr>
<td>Canis familiaris</td>
<td>canFam1</td>
<td>Yes</td>
<td>Drosophila melanogaster</td>
<td>dm2</td>
<td>No</td>
</tr>
<tr>
<td>Bos taurus</td>
<td>bosTaur1</td>
<td>Yes</td>
<td>Anopheles gambiae</td>
<td>anoGam1</td>
<td>No</td>
</tr>
<tr>
<td>Monodelphis domestica</td>
<td>monDom1</td>
<td>Yes</td>
<td>Caenorhabditis elegans</td>
<td>ce2</td>
<td>No</td>
</tr>
<tr>
<td>Gallus gallus</td>
<td>galGal2</td>
<td>Yes</td>
<td>Saccharomyces cerevisiae</td>
<td>sacCer1</td>
<td>No</td>
</tr>
<tr>
<td>Xenopus tropicalis</td>
<td>xenTro1</td>
<td>Yes</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

✓ 100 diverged copies in a Gigabase
✓✓ 60 highly similar copies in a Megabase

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

Human Genome: $3 \times 10^9$ letters

1.5% known function

>50% junk

http://bejerano.stanford.edu
The LF SINE (for Lobefin Fish / “Living Fossil”)

A Box & B Box
Rat Serine tRNA
Coelacanth SINE

not similar to any known repeat
polyA

Reconstruction

target site
duplications
>360My Old and Going Strong

Upto 80% id between Coelacanath SINE and some human instances, inc uc.338.

<table>
<thead>
<tr>
<th>Species</th>
<th>UCSC Assembly</th>
<th>LF-SINE Detected</th>
<th>Species</th>
<th>UCSC Assembly</th>
<th>LF-SINE Detected</th>
</tr>
</thead>
<tbody>
<tr>
<td>Homo sapiens</td>
<td>hg17</td>
<td>Yes</td>
<td>Danio rerio</td>
<td>danRe2</td>
<td>No</td>
</tr>
<tr>
<td>Pan troglodytes</td>
<td>panTro1</td>
<td>Yes</td>
<td>Tetraodon nigroviridis</td>
<td>tetNig1</td>
<td>No</td>
</tr>
<tr>
<td>Macaca mulatta</td>
<td>rheMac1</td>
<td>Yes</td>
<td>Takifugu rubripes</td>
<td>frl</td>
<td>No</td>
</tr>
<tr>
<td>Mus musculus</td>
<td>mm6</td>
<td>Yes</td>
<td>Ciona intestinalis</td>
<td>cil</td>
<td>No</td>
</tr>
<tr>
<td>Rattus norvegicus</td>
<td>rn3</td>
<td>Yes</td>
<td>Strongylocentrotus purpuratus</td>
<td>strPur1</td>
<td>No</td>
</tr>
<tr>
<td>Canis familiaris</td>
<td>canFam1</td>
<td>Yes</td>
<td>Drosophila melanogaster</td>
<td>dm2</td>
<td>No</td>
</tr>
<tr>
<td>Bos taurus</td>
<td>bosTaur1</td>
<td>Yes</td>
<td>Anopheles gambiae</td>
<td>anoGam1</td>
<td>No</td>
</tr>
<tr>
<td>Monodelphis domestica</td>
<td>monDom1</td>
<td>Yes</td>
<td>Caenorhabditis elegans</td>
<td>ce2</td>
<td>No</td>
</tr>
<tr>
<td>Gallus gallus</td>
<td>galGal2</td>
<td>Yes</td>
<td>Saccharomyces cerevisiae</td>
<td>sacCer1</td>
<td>No</td>
</tr>
<tr>
<td>Xenopus tropicalis</td>
<td>xenTro1</td>
<td>Yes</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
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

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

http://bejerano.stanford.edu

[Bejerano et al., Nature 2006] 29
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</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>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>210</td>
</tr>
<tr>
<td><em>Macaca mulatta</em></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>235</td>
</tr>
<tr>
<td><em>Bos taurus</em></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>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>87</td>
</tr>
<tr>
<td><em>Monodelphis domestica</em></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>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>10</td>
<td>14</td>
<td>26</td>
</tr>
</tbody>
</table>

Human instances cluster together, found <1Mb from 35 TFs (P<3*10^{-6}).
ISL1 is a neuro-developmental gene, also expressed in testis. Three previously known enhancers are conserved across vertebrates.
Repeat made Regulatory Region

Conserved Element \( \rightarrow \) Minimal Promoter \( \rightarrow \) Reporter Gene

in situ

transgenic

http://bejerano.stanford.edu
Co-option into Different Roles
The Co-Optionome

quantify co-option

transposition event

functional elements

LF-SINE, DeuSINE, MER121, ...

[Lowe, Bejerano & Haussler, Submitted]
From junk DNA to pathway recruitments?

1. A portion of the genome containing a new solitary replication: 

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]
Bejerano Lab: Marry Development & Genomics

Origins & Evolution  Functions & Encoding  Contribution to Human Disease

http://bejerano.stanford.edu
Kudos

UC Santa Cruz: David Haussler, Sofie Salama, Jim Kent, Craig Lowe, Sol Katzman, Andy Kern, *Everyone*…
Berkeley Labs (LBNL): Eddy Rubin, Nadav Ahituv

Stanford faculty, and Bejerano Lab:

Cory McLean  
Sarah Aerni  
Phil Lacroute  
Shoa Clarke  
Abe Bassan  
Marina Sirota  

Gill Bejerano  
bejerano@stanford.edu

http://bejerano.stanford.edu