Global Genome Initiative

Preserving and Understanding the Genomic Diversity of Life

part of the Smithsonian Institute for Biodiversity Genomics
Smithsonian Institution

- Largest museum (19) & research complex (9), citizen owned and supported
- Encyclopedia of Life (EOL)
- Consortium for the Barcode of Life (CBOL)
- Integrated Taxonomic Information System (ITIS)
- Moorea BIOCODE
- Smithsonian Global Earth Observatories (SIGEO)
- Marine Global Observatories MarineGEO
- Biodiversity Heritage library (BHL)
National Museum of Natural History

- 200 PhD’s (staff and affiliated)
- 127 million specimens and artifacts
- Two campuses: Mall (1M ft²) and Museum Support Center in Suitland, MD (415k ft²)
~5-7M visitors per year
Biodiversity Genomics: costs and progress
<table>
<thead>
<tr>
<th>Before</th>
<th>After</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hard-to-find, ambiguous quality tissues ambiguously owned by individual PI’s</td>
<td>Publically accessible, genome-quality tissues in enterprise biorepositories following best practices and Int. treaties</td>
</tr>
<tr>
<td>Expensive “boutique” sequencing of a few model genomes</td>
<td>Affordable, coordinated, sequencing of a thoughtful synopsis of all of Life</td>
</tr>
<tr>
<td>Phenotype-based, expert- limited taxonomy as well as Environmental biology, evolution, conservation, ecology, biotech</td>
<td>Approximate IDs of most organisms anywhere</td>
</tr>
<tr>
<td></td>
<td>Precise, scalable, cheap tools</td>
</tr>
</tbody>
</table>
GGI Research Questions

• Origins, Diversity, and Patterns of Life
  – What is Earth’s true “Tree of Life”? 
  – How many kinds of life exist? 
  – What key innovations drove evolutionary success and how did they evolve?

• Adaptation and Evolutionary Novelty
  – What makes individual organisms resilient to change? 
  – Genomic tools to monitor and mediate biological and environmental change? 
  – Occurrence and dynamics of diseases in wild species that threaten humans?
Success (Phase I)

Preservation of 50% of the major branches of Life in 6 years (~10,000 families, 80-100,000 genera)

- Research impact and productivity
- Partnerships and networks
- Near-Universal IDs
- Global networked biorepositories
- Genome quality samples
- Public understanding, literacy
Feasibility: Phylogeny

- **DOMAINS**: 3
- **PHYLA / DIVISIONS**: 91
- **CLASSES**: 364
- **ORDERS**: 1413
- **FAMILIES**: ~9642\(^1\)
- **“GENERA”**: ~200,000\(^2\)
- **SPECIES**: >15,000,000

\(^1\)6537, \(^2\)52,169 in Genbank (03/2011)
“Families” Discovery

**Families**

- Bees
- Mammals
- Birds
- Spiders
- Angiosperms
- Bacteria-Archaea

Earliest Description:

- 1750
- 1770
- 1810
- 1830
- 1850
- 1870
- 1890
- 1910
- 1930
- 1950
- 1970
- 1990
- 2010

**Percent 2010 Total**

- 0.00
- 0.10
- 0.20
- 0.30
- 0.40
- 0.50
- 0.60
- 0.70
- 0.80
- 0.90
- 1.00
Hyalogyrinidae
Hyalogyra expansa B. A. Marshall, 1988

Sapayoaidae
(Sapayoa aenigma Hunt 1903)

Limnognathidae
(Limnognathia maerski Kristensen & Funch 2000)

Craseonycteridae
(Craseonycteris thonglongyai Hill, 1974)

Godzilliidae
(Godzillius robustus Schram et al., 1986)

Protanguillidae
206 Images

GGI Knowledge Portal “Oaks”

756 barcodes

7572 tissues

58,460 sequences

ALF “All Living Families

1,017,951 localities
Family Fagaceae

Beech family

Eukaryota > Fagales

Ref: Stevens, 2013

Source:

<table>
<thead>
<tr>
<th>Source</th>
<th>Count</th>
<th>% Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>BHL</td>
<td>6672</td>
<td>100</td>
</tr>
<tr>
<td>BOLD</td>
<td>756</td>
<td>98</td>
</tr>
<tr>
<td>EOL</td>
<td>208</td>
<td>100</td>
</tr>
<tr>
<td>GBIF</td>
<td>1017951</td>
<td>100</td>
</tr>
<tr>
<td>GenBank sequences</td>
<td>58460</td>
<td>98</td>
</tr>
<tr>
<td>GGBN</td>
<td>7572</td>
<td>100</td>
</tr>
</tbody>
</table>

GGI score: 99

File: Quercus petraea foliage acorns Bulgaria.jpg
By Nenko Lazarov via Wikimedia Commons

http://ggi.eol.org/
## All Living Families

Basic statistics on six data sources for the 9,642 families of life on Earth. All values are counts.

<table>
<thead>
<tr>
<th></th>
<th>Min</th>
<th>Max</th>
<th>Mean</th>
<th>Median</th>
<th>StDev</th>
<th>Interquartile</th>
<th>Zero Count</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>BHL</td>
<td>0</td>
<td>11,246,645</td>
<td>11,804</td>
<td>17</td>
<td>198,807</td>
<td>199</td>
<td>2,698</td>
<td>2,795,476</td>
</tr>
<tr>
<td>BOLD</td>
<td>0</td>
<td>3,175</td>
<td>11</td>
<td>-</td>
<td>77</td>
<td>2</td>
<td>5,703</td>
<td>1,161,799</td>
</tr>
<tr>
<td>EOL</td>
<td>0</td>
<td>26,435</td>
<td>290</td>
<td>17</td>
<td>1,160</td>
<td>120</td>
<td>2,372</td>
<td>106,047</td>
</tr>
<tr>
<td>GBIF</td>
<td>0</td>
<td>20,388,917</td>
<td>43,590</td>
<td>245</td>
<td>444,943</td>
<td>3,240</td>
<td>1,383</td>
<td>420,646,292</td>
</tr>
<tr>
<td>GGBN</td>
<td>0</td>
<td>12,141</td>
<td>7</td>
<td>-</td>
<td>168</td>
<td>-</td>
<td>8,515</td>
<td>77,314</td>
</tr>
<tr>
<td>NCBI</td>
<td>0</td>
<td>39,785</td>
<td>120</td>
<td>-</td>
<td>1,052</td>
<td>16</td>
<td>5,036</td>
<td>113,911,762</td>
</tr>
<tr>
<td>Total</td>
<td>0</td>
<td>31,717,098</td>
<td>55,823</td>
<td>279</td>
<td>646,208</td>
<td>3,577</td>
<td>25,707</td>
<td>538,702,340</td>
</tr>
</tbody>
</table>

1 Ruggiero (2014, 2015)
• Micrognathozoa: *Limnognathia maerski* monotypic Class discovered in 1994 on Disko Island, Greenland, 0.1mm.

• Only TWO DATA POINTS

<table>
<thead>
<tr>
<th>Source</th>
<th>Count</th>
<th>% score</th>
</tr>
</thead>
<tbody>
<tr>
<td>BHL</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>BOLD</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>EOL</td>
<td>1</td>
<td>60</td>
</tr>
<tr>
<td>GBIF</td>
<td>1</td>
<td>15</td>
</tr>
<tr>
<td>GenBank</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>sequences</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>GGBN</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>GGI score</td>
<td>13</td>
<td></td>
</tr>
</tbody>
</table>
Feasibility: Taxonomy

Barcode (COI) ID “radius” (e.g., European spiders)

- 50 families, 313 genera, and 821 species
- 873 sequences blasted against themselves
- 91% correct at family level\(^1\)
- 85% correct at genus level\(^1\)

\(^1\) PIdent > 0.95
Feasibility: Moorea Biocode

Phyla: 74%
Class: 61%
Order: 42%
Family: 23%
Feasibility: Forest Global Earth Observatories

Smithsonian Institution Forest Earth Observatory

40 plots, 10,500 species, 4,346 genera ("trees")
~60% world total?
Ocean Sampling Day and Marine GEO Collaborators

28 Marine Genomic Observatories
<table>
<thead>
<tr>
<th></th>
<th>Chesapeake Bay</th>
<th>Fort Pierce FL</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sequences</td>
<td>572,290</td>
<td>409,613</td>
</tr>
<tr>
<td>Species</td>
<td>1,204</td>
<td>1,391</td>
</tr>
<tr>
<td>Unknown</td>
<td>40.9</td>
<td>28.3</td>
</tr>
</tbody>
</table>

...91% correct for families; 85% for genera
NGS sequencing of 16S and 18S prokaryotes and eukaryotes

Emphasis on invasive parasites, pathogens, and toxic microorganisms in the ballast water of cargo ships

PI: Greg Ruiz (SERC)
SI Co-PIs: Rob Fleischer (CCEG-SCBI), Katrina Lohan (SERC, CCEG-SCBI)

Some toxic dinoflagellates cause harmful algal blooms that are detrimental to humans and ecosystems.

E. coli has been found in ballast water.

Can the global distribution of *Perkinsus olseni*, a major parasite of bivalves, be attributed to transport in ballast water?
Laboratories of Analytical Biology
NMNH Biorepository
4-5M 2ml tube capacity

58 Freezers

24 Nitrogen Tanks
A Collaborative Effort to Cryo-preserve and Provide Access to Genomic Samples from Across the Tree of Life

The Global Genome Biodiversity Network (GGBN) is a global network of well-managed collections of genomic tissue samples from across the Tree of Life, benefiting society through biodiversity research, development and conservation. This network will foster collaborations among repositories of molecular biodiversity in order to ensure quality standards, improve best practices, secure interoperability, and harmonize exchange of material in accordance with national and international legislation and conventions.

**Goals**

- Provide genome-quality samples from across the Tree of Life for research, training, and development, thereby contributing to the conservation of global genetic diversity for generations to come.
- Provide open access to a global data management system hosting the aggregated primary specimen data and metadata for all the member institutions.

**GGBN Counts**

- Samples Online = 100,471
- Taxa Online = 23,349
- Collections Online = 12

**News & Updates**

**A new version of the GGBN portal is currently under construction, with a beta version available at data.ggbn.org/ggbn_new. Final release of new portal is scheduled for December 2015. Various pages or links may have different designs while the transition is underway.**
GGBN Value Proposition One

- Make genomic collections *discoverable* for research through a networked community of biorepositories
  - Provide standardized tool to make collections discoverable;
  - Provide community standards and best practices curation and management of samples and data
  - Build international community for support and access to storage facilities for and access to information on their genomic collections.
GGBN Value Proposition Two

• Provide trusted and transparent access to genomic samples for users and contributors through an ABS framework—supporting the trust between biodiversity-rich countries and organizations.

• Provides strategic guidance for growth of global biorepositories to preserve Life’s genome.
GGBN Data Portal Milestones

Oct 2011  1st Intl. workshop, DC
July 2012  1st Intl. meeting, Copenhagen
2013     BGI, China National Gene Bank, Ocean Sampling Day
Spring 2013 NSF Workshop (USA)
August 2013: GGBN & DNA Bank Network sites merged
Fall 2013  DFG Grant (3 FTE’s 2 years)
November 2014: Private beta release of GGBN Data Portal
June 2015: 1st public release of GGBN Data Portal
November 2015: Submit GGBN Data Standard to TDWG review and ratification process
November 2015: 2nd public release of GGBN Data Portal
June 2016  2nd Intl Meeting, Berlin
Observer Membership

Supports GGBN Goals, Ethics, Standards, Practices, access to all support and tools, works to Associate Mem.

Associate Membership

Institutional Biorepositories and:

- Preserves genomic collections and metadata
- Registers as a biorepository (GRBio);
- Progress toward enterprise-level, reliable data system;
- Will contribute GGBN standard data

Core Membership

- Contributing data to GGBN.
Today: 36 GGBN members worldwide

172K Samples
23K Taxa
12 Collections

16% Families in GGBN
8% Genera in GGBN
Nagoya Protocol on Access to Genetic Resources and the Fair and Equitable Sharing of Benefits

(Convention on Biological Diversity)

- Oct 2014
- Greater legal certainty and transparency
- Promote and encourage research
- Awareness-raising & technology transfer
- In-country research capability and institutions
Feasibility: GGI Gardens!

DC Area
3 Gardens
>50% plant families
1000’s genera

Expanding to
Berlin
Copenhagen
Memphis
Chicago
Australia
Imagine that ... We understand the relationships among all living things.

We preserve the genomes of all major branches of the tree of life.

We enable robotic, industrial, identification of most organisms.

We apply the full range of technologies from nature to solve human problems.
Only with Friends!
Thanks!

FAPESP
UNICAMP
Glynn
Derlene
e vocês
<table>
<thead>
<tr>
<th>Column</th>
<th>&gt; 9416 bp</th>
<th>&lt; 9416 bp</th>
<th>total</th>
<th>%&gt;9416 bp</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>3503</td>
<td>61565</td>
<td>65070</td>
<td>6%</td>
</tr>
<tr>
<td>3</td>
<td>5026</td>
<td>79952</td>
<td>84982</td>
<td>6%</td>
</tr>
<tr>
<td>4</td>
<td>12537</td>
<td>27177</td>
<td>39718</td>
<td>32%</td>
</tr>
<tr>
<td>5</td>
<td>5965</td>
<td>8789</td>
<td>14759</td>
<td>40%</td>
</tr>
<tr>
<td>6</td>
<td>8906</td>
<td>6386</td>
<td>15298</td>
<td>58%</td>
</tr>
<tr>
<td>7</td>
<td>5767</td>
<td>2612</td>
<td>8387</td>
<td>69%</td>
</tr>
<tr>
<td>8</td>
<td>1096</td>
<td>29235</td>
<td>30339</td>
<td>4%</td>
</tr>
<tr>
<td>9</td>
<td>4876</td>
<td>28195</td>
<td>33081</td>
<td>15%</td>
</tr>
</tbody>
</table>
