Genomics in Brazil

Andrew J.G. Simpson, Ph.D.
Ludwig Institute for Cancer Research
Creation of ONSA in 1997:
The beginning of Large-scale genome sequencing in Brazil

- Virtual Institute
- Sponsored by FAPESP. The State of São Paulo Science Foundation
Creation of ONSA

- Boost Brazilian competence in the emerging area of Genomics
- Promote scientific collaboration
- Study and organism of significant economic impact in the State of São Paulo
How the network functions

Coordinating lab

Preparation of libraries

Sequencing labs

Database

annotation

Bioinformatics lab
Why Xylella fastidiosa?

- Citrus industry in Brazil: US$ 2 billion/yr
- 90% located in the State of São Paulo
- 400,000 employments
- The disease has significant economic impact: US$ 100 million
- First plant pathogen to be sequenced.
NO Avr or Hrp genes!
Citrus pathogen sequenced

Isotope geology
Strange sulphates

AIDS
Mbeki responds to critics

Molecular logic
Chemistry meets computing

nature jobs
focus on biochemistry
The Xylella project started the genomics era

April 24, 2001

Model for Research Rises in a Third World City

By LARRY ROHTER

ÃO PAULO, Brazil — It has no laboratories or research teams of its own, only a modest administrative staff working out of a nondescript building in a residential neighborhood here. But through canny management and careful choices, the Research Support Foundation of the State of São Paulo is rapidly becoming a powerhouse in genomics and a model for scientific investigation in the third world...
SAMBA, football and...genomics. The list of things for which Brazil is renowned has suddenly got longer. Only a few days after publishing, on July 13th, the first-ever sequence of the genome of a plant pathogen, scientists at Sao Paulo’s state research agency, Fapesp, were due to announce, on July 21st, another success—the composition of 279,000 human expressed-sequence tags, small pieces of DNA that allow genes to be located along chromosomes.
X. citri
X. campestris

Xylella strains
Leifsonia
Eucalyptus
Human pathogen
Human Cancer Genome Project
<table>
<thead>
<tr>
<th>Category</th>
<th>Count</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>rRNA</td>
<td>45297</td>
<td>3.84%</td>
</tr>
<tr>
<td>Bacteria</td>
<td>53905</td>
<td>4.57%</td>
</tr>
<tr>
<td>Known Human Genes</td>
<td>206429</td>
<td>17.49%</td>
</tr>
<tr>
<td>Unigene Contigs</td>
<td>268203</td>
<td>22.73%</td>
</tr>
<tr>
<td>Non-unigene ESTs</td>
<td>97094</td>
<td>8.23%</td>
</tr>
<tr>
<td>Paralogs</td>
<td>8071</td>
<td>0.68%</td>
</tr>
<tr>
<td>Non-Human Protein</td>
<td>11156</td>
<td>0.95%</td>
</tr>
<tr>
<td>ESTs</td>
<td>32</td>
<td>0.00%</td>
</tr>
<tr>
<td>DNA</td>
<td>126</td>
<td>0.01%</td>
</tr>
<tr>
<td>Repeats</td>
<td>76726</td>
<td>6.50%</td>
</tr>
<tr>
<td>No matches</td>
<td>349598</td>
<td>29.62%</td>
</tr>
<tr>
<td>Human Protein</td>
<td>991</td>
<td>0.08%</td>
</tr>
<tr>
<td>TOTAL</td>
<td>1180149</td>
<td></td>
</tr>
</tbody>
</table>
ORESTES Utilization

TIGR Human Gene Index

Attributions

A significant number of ESTs used to construct this index were generated by:

- Washington University School of Medicine, Genome Sequencing Center
- The Institute for Genomic Research
- Genetics Human Genome Research Center
- University of Toronto
- Marsh Information Center for Protein Sequence
Shotgun sequencing of the human transcriptome with ORF expressed sequence tags

The contribution of 700,000 ORF sequence tags to the definition of the human transcriptome


Ludwig Institute for Cancer Research, (LICR), Ribeirão Preto, Brazil. Department of Genetics, Federal University of São Paulo, UNIFESP, 04109-032, São Paulo, Brazil. Department of Biophysics and Molecular Biology, Institute of Biophysics, Federal University of São Paulo, 05508-090, São Paulo, Brazil. Department of Medical Sciences, Institute of Biophysics, Federal University of São Paulo, 05508-090, São Paulo, Brazil. Department of Molecular Biology, Federal University of São Paulo, UNIFESP, 04109-032, São Paulo, Brazil. Department of Genetics, Federal University of São Paulo, 05508-090, São Paulo, Brazil. Department of Biophysics and Molecular Biology, Institute of Biophysics, Federal University of São Paulo, 05508-090, São Paulo, Brazil. Department of Medical Sciences, Institute of Biophysics, Federal University of São Paulo, 05508-090, São Paulo, Brazil. Department of Genetics, Federal University of São Paulo, 05508-090, São Paulo, Brazil. Department of Medical Sciences, Institute of Biophysics, Federal University of São Paulo, 05508-090, São Paulo, Brazil.

Open reading frame expressed sequence tags (ORFEST) differ from conventional ESTs by providing sequence data from the central coding portion of transcripts. We generated a total of 66,635 ORFEST sequences from 28 human tissues and used a subset of the data that correspond to a set of 35,085 full-length mRNAs as a means of assessing the efficiency of this strategy and its potential contribution to the definition of the human transcriptome. We estimate that ORFESTs constitute 80% of all highly and moderately expressed genes, and 50% of all expressed human genes. Moreover, 50% of all expressed human genes are expressed as ORFESTs and 40% of those genes expressed as ORFESTs are expressed as full-length mRNAs. Our results suggest that ORFESTs offer a new perspective on the human transcriptome, potentially providing new insights into the biology of human disease. The data also suggest that ORFESTs may provide a useful alternative to full-length cDNA cloning.

The identification of all human genes and transcripts remains a goal of highest priority and a rate-limiting step in progress toward the completion of the human genome sequence. The complexity and variability of human gene structure present a direct challenge to genome sequencing and the development of databases. However, a significant number of genes are expressed as ORFESTs and may be more accurately represented by ORFESTs than by full-length cDNAs. These data suggest that ORFESTs offer a new perspective on the human transcriptome and may provide a useful alternative to full-length cDNA cloning.
The Federal Government decided to expand the genome project at the national level and launched the **Brazilian Genome Project** (Dec. 2000)

- Part of the National Program of Biotechnology and Genetic Resources (R$250 m)
- Comprises a network of 25 sequencing laboratories
- Objective: to sequence the genome of *Chomobacterium violaceum*
  - abundant in the waters of the Amazon Region
  - opportunistic human pathogen (infection is fatal)
  - produces violacein of trypanocidal and antibiotic activity
  - Produces a polyester similar to propylene and polyethylene
Organism choice

*Chromobacterium violaceum* is a free-living organism and have a genome size in the range of 4.5 to 4.7 Mbp.
Europe
9.910.000 km²

Brazil
8.512.000 km²

~4300 km
Regional Genome Projects

Human Health

Rede do Nordeste
Leishmania chagasi

Rede do Centro-oeste
Paracocidioides brasiliensis

Rede de Minas Gerais
Schistosoma mansoni
Regional Genome Projects

- **Agriculture**

  - **Rede Bahia/ Unicamp**
    - *Crinipellis perniciosa*
  
  - **Rede Riogene**
    - *Gluconacetobacter diazotrophicus*
  
  - **Rede Genopar**
    - *Herbaspirillum seropedicae*
Genomics Networks

- Are rapid and cheap to create
- Permit cutting edge science
- Create a critical mass
- Overcome geographic isolation
- Foster a collaborative spirit
- Enable developing countries to compete
- Permit scientist to scientist interaction
- Are easily abandoned
- Stimulate local high quality research