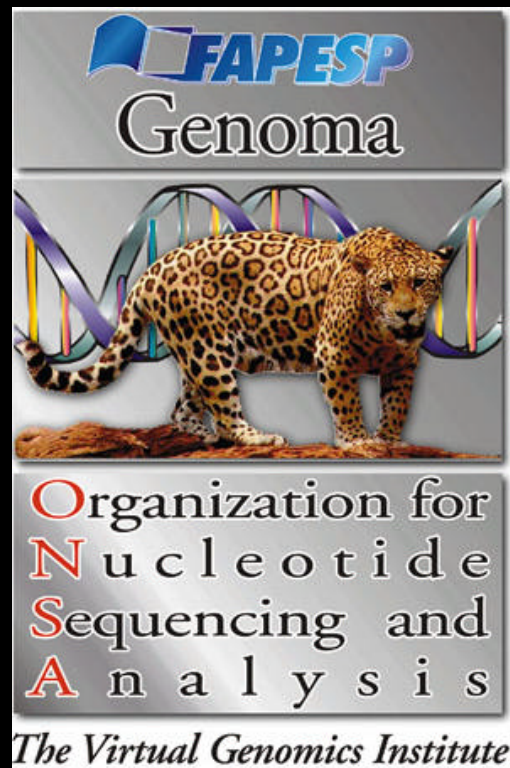


# Genomics in Brazil

Andrew J.G. Simpson, Ph.D.  
Ludwig Institute for Cancer  
Research

## Creation of ONSA in 1997:

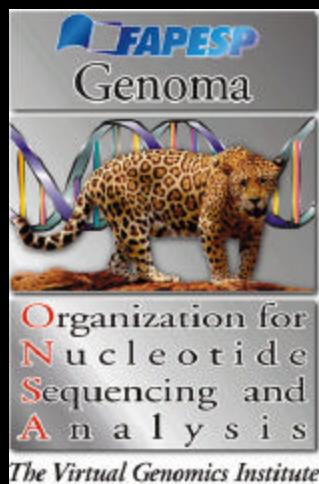
The beginning of Large-scale genome sequencing in Brazil



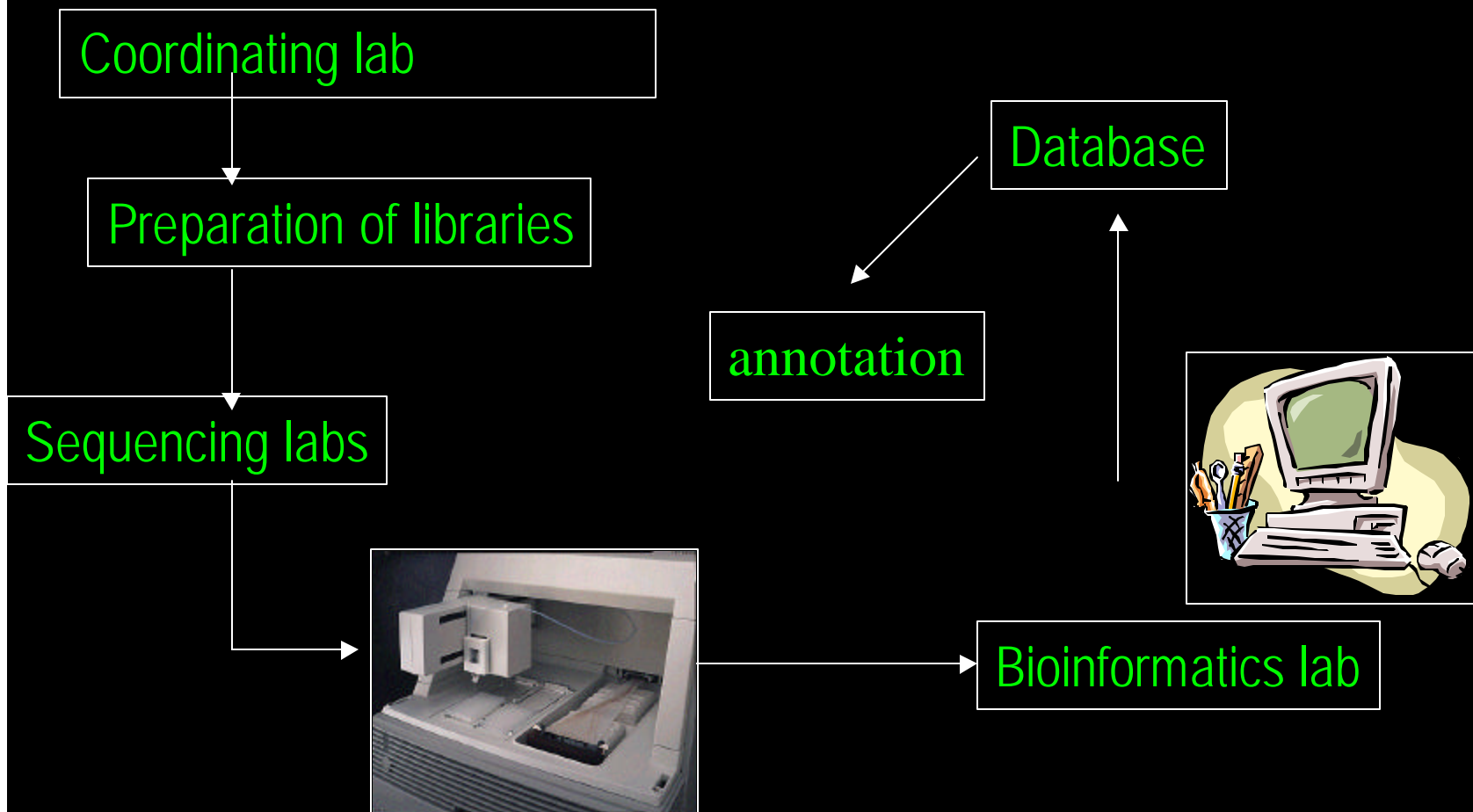
- Virtual Institute
- Sponsored by FAPESP. The State of São Paulo Science Foundation

## 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

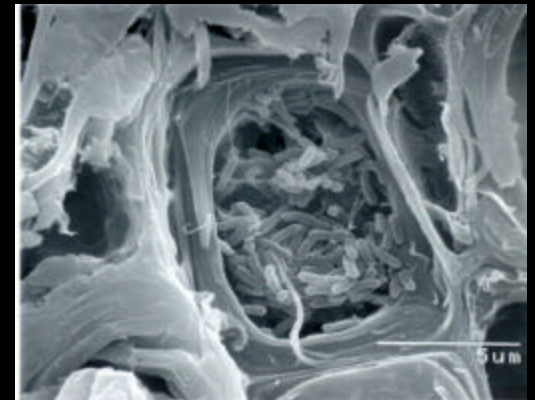


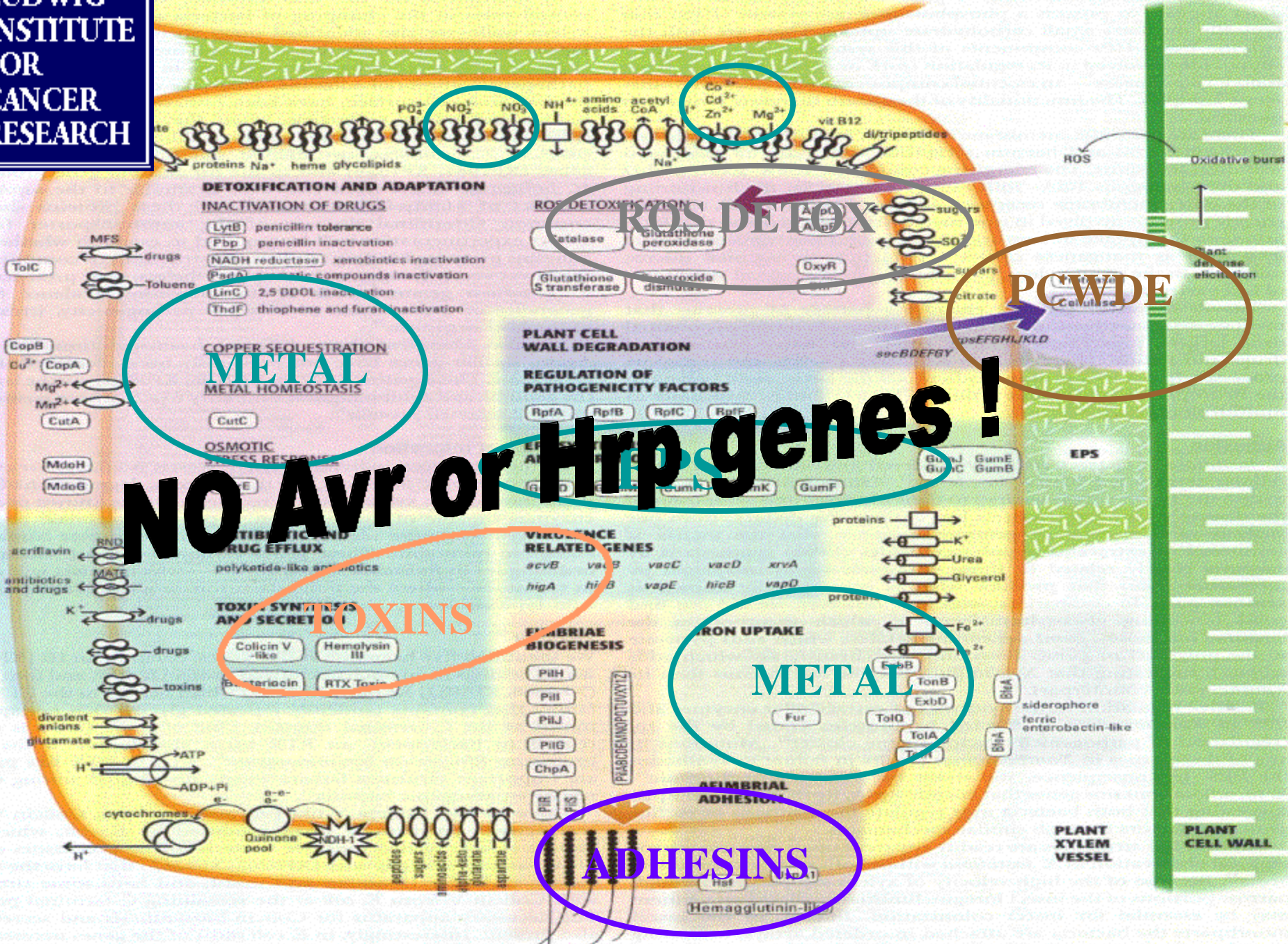
# How the network functions



## 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.





LUDWIG  
INSTITUTE  
FOR  
CANCER  
RESEARCH

13 July 2000

International weekly journal of science

# nature

£10.00

www.nature.com

## 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...

## The Economist

July 22nd-29th, 2000 - Ed. no. 8180

Brazilian science Fruits of co-operation

Peter Collins

S A O P A U L O

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.

## ONSA Today

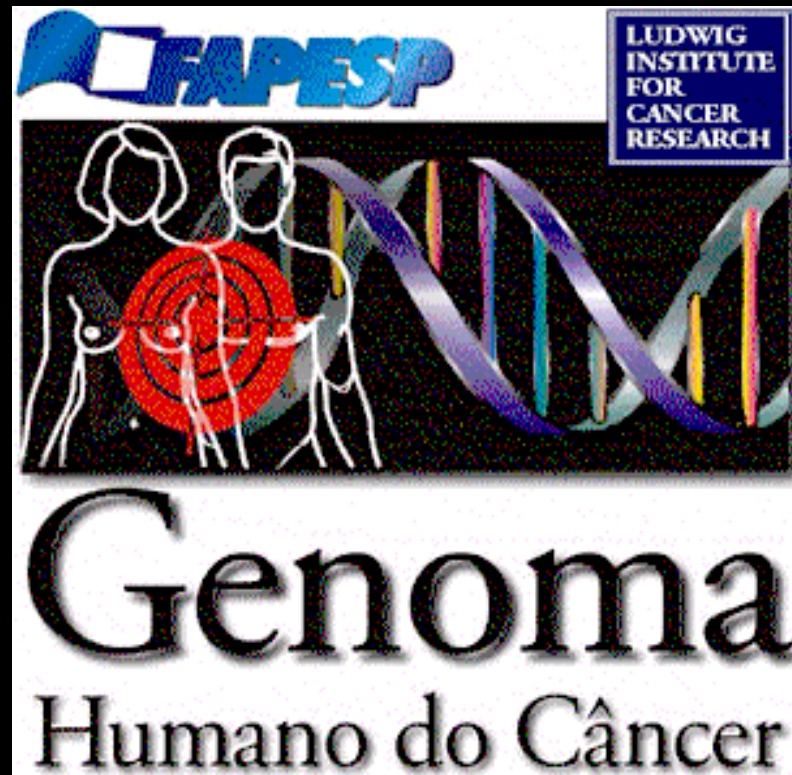


*X. citri*  
*X. campestris*

*Xylella strains*  
*Leifsonia*  
*Eucalyptus*  
*Human pathogen*

LUDWIG  
INSTITUTE  
FOR  
CANCER  
RESEARCH

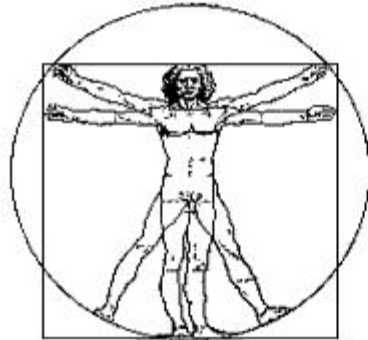
# Human Cancer Genome Project



# Project Productivity

rRNA	45297 ( 3.84%)
Bacteria	53905 ( 4.57%)
Known Human Genes	206429 ( 17.49%)
Unigene Contigs	268203 ( 22.73%)
Non-unigene ESTs	97094 ( 8.23%)
Paralogs	8071 ( 0.68%)
Non-Human Protein	11156 ( 0.95%)
ESTs	32 ( 0.00%)
DNA	126 ( 0.01%)
Repeats	76726 ( 6.50%)
No matches	349598 ( 29.62%)
Human Protein	991 ( 0.08%)
TOTAL	1180149

# 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.](#)



[FAPESP/LICR-Human Cancer Genome Project.](#)



[The Institute for Genomic Research.](#)



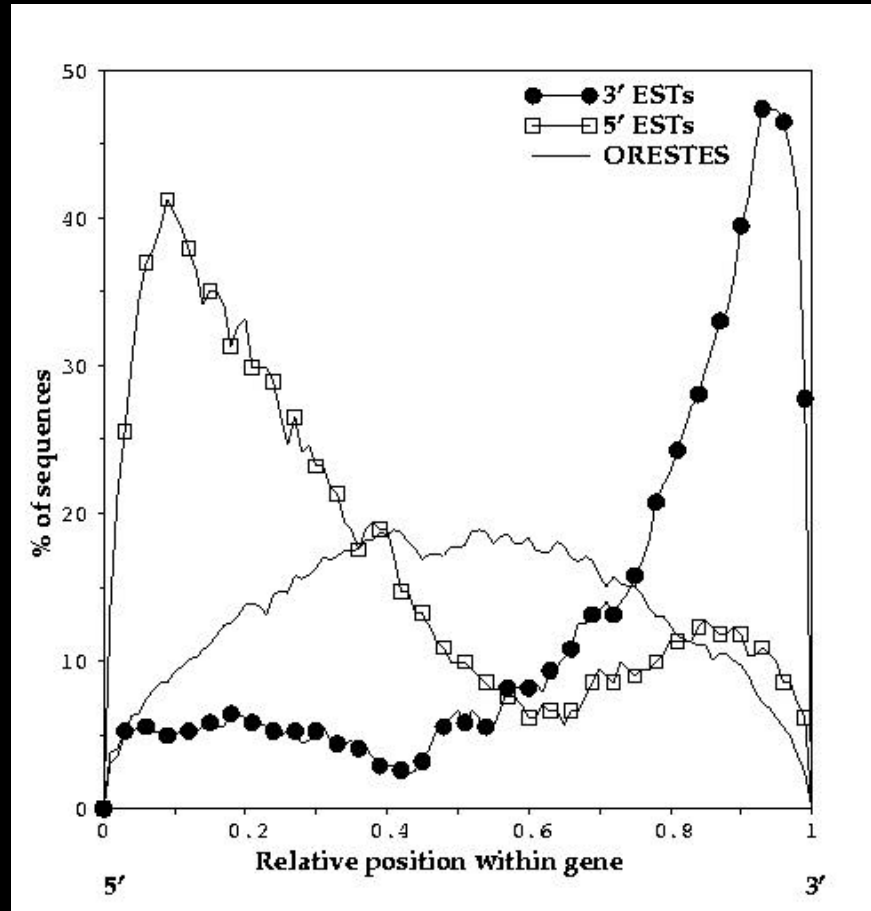
[Genethon Human Genome Research Center.](#)



[University of Toronto.](#)



[Munich Information Center for Protein Sequence.](#)



## Shotgun sequencing of the human transcriptome with ORF expressed sequence tags

Emmanuel Dias Neto<sup>a</sup>, Ricardo Garcia Correa<sup>a</sup>, Sergio Verjovski-Almeida<sup>b</sup>, Marcelo R. S. Briones<sup>c</sup>, Maria Aparecida Nagai<sup>d</sup>, Wilson da Silva, Jr.<sup>e</sup>, Marco Antonio Zago<sup>e</sup>, Silvana Bordin<sup>f</sup>, Fernando Ferreira Costa<sup>f</sup>, Gustavo Henrique Goldman<sup>g</sup>, Alex F. Carvalho<sup>a</sup>, Adriana Matsukuma<sup>b</sup>, Gilson S. Baia<sup>b</sup>, David H. Simpson<sup>h</sup>, Adriana Brunstein<sup>a</sup>, Paulo S. L. de Oliveira<sup>a</sup>, Philipp Bucher<sup>i</sup>, C. Victor Jongeneel<sup>j</sup>, Michael J. O'Hare<sup>k</sup>, Fernando Soares<sup>l</sup>, Ricardo R. Brentani<sup>a</sup>, Luis F. L. Reis<sup>a</sup>, Sandro J. de Souza<sup>a</sup>, and Andrew J. G. Simpson<sup>a,m</sup>

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

Anamaria A. Camargo<sup>1</sup>, Helena P. B. Samola<sup>2</sup>, Emmanuel Dias-Neto<sup>3</sup>, Daniel F. Simão<sup>4</sup>, Italo A. Milgotto<sup>5</sup>, Marcelo R. S. Briones<sup>6</sup>, Fernando F. Costa<sup>7</sup>, Maria A.parecida Nagai<sup>8</sup>, Sergio Verjovski-Almeida<sup>9</sup>, Marco A. Zago<sup>1</sup>, Luis Eduardo C. Andrade<sup>10</sup>, Helaine Carneiro<sup>11</sup>, Hamza F. A. El-Dorry<sup>12</sup>, Enliza M. Espreafico<sup>13</sup>, Angelita Haber-Gamal<sup>14</sup>, Daniel Giannella-Neto<sup>15</sup>, Gustavo H. Goldman<sup>16</sup>, Arthur Gruber<sup>17</sup>, Christine Hackel<sup>18</sup>, Edna T. Kimura<sup>19</sup>, Rui M. B. Maciel<sup>20</sup>, Susly K. N. Marie<sup>21</sup>, Elizabeth A. L. Martins<sup>22</sup>, Marina P. Nóbrega<sup>23</sup>, Maria Lúcia Paço-Larson<sup>24</sup>, Maria Inês M. C. Pardini<sup>25</sup>, Gonçalves G. Pereira<sup>26</sup>, João Bosco Pesquero<sup>27</sup>, Vanderlei Rodrigues<sup>28</sup>, Sílvia R. Rogatto<sup>29</sup>, Ismael D. C. G. da Silva<sup>30</sup>, Mari C. Sogayar<sup>31</sup>, Maria de Fátima Souza<sup>32</sup>, Bolza H. Tajara<sup>33</sup>, Sandro R. Valentini<sup>34</sup>, Fernando L. Aliberto<sup>35</sup>, Maria Elisabete J. Amara<sup>36</sup>, Ivy Azevedo<sup>37</sup>, Ullone A. T. Arnaldi<sup>38</sup>, Angela M. de Assis<sup>39</sup>, Mário Henrique Bangtson<sup>40</sup>, Nadia Aparecida Bergamo<sup>41</sup>, Vanessa Bombonato<sup>42</sup>, Maria E. R. de Camargo<sup>43</sup>, Renata A. Canavaroli<sup>44</sup>, Dirco M. Carraro<sup>45</sup>, Janete M. Coutinho<sup>46</sup>, Maria Lúcia C. Corrêa<sup>47</sup>, Rosana F. R. Corrêa<sup>48</sup>, Maria Cristina R. Costa<sup>49</sup>, Cynthia Curdo<sup>50</sup>, Paula O. M. Hokama<sup>51</sup>, Ari J. S. Ferreira<sup>52</sup>, Gilberto K. Furuzawa<sup>53</sup>, Taiseko Gushiken<sup>54</sup>, Paulo L. Ho<sup>55</sup>, Elza Kimura<sup>56</sup>, José E. Krieger<sup>57</sup>, Luciana C. C. Leite<sup>58</sup>, Paromita Majumder<sup>59</sup>, Mozart Martins<sup>60</sup>, Everaldo R. Marques<sup>61</sup>, Anely S. A. Melo<sup>62</sup>, Monica Melo<sup>63</sup>, Carlos Alberto Mestrina<sup>64</sup>, Elisabete C. Miracca<sup>65</sup>, Daniela C. Miranda<sup>66</sup>, Ana Lucia T. O. Nascimento<sup>67</sup>, Francisco G. Nóbrega<sup>68</sup>, Elda P. B. Ojopi<sup>69</sup>, José Rodrigo C. Pandolfi<sup>70</sup>, Ludana G. Passos<sup>71</sup>, Aline C. Pravedo<sup>72</sup>, Paula Rahal<sup>73</sup>, Claudio A. Rainho<sup>74</sup>, Eduardo M. R. Reis<sup>75</sup>, Marcelo L. Ribeiro<sup>76</sup>, Nancy da Rosa<sup>77</sup>, Renata G. de Sá<sup>78</sup>, Magaly M. Sales<sup>79</sup>, Simone Cristina Sant'anna<sup>80</sup>, Mariana L. dos Santos<sup>81</sup>, Aline M. da Silva<sup>82</sup>, Neusa P. da Silva<sup>83</sup>, Wilson A. Silva, Jr.<sup>84</sup>, Rosana A. da Silveira<sup>85</sup>, Josane F. Sousa<sup>86</sup>, Daniella Stecco<sup>87</sup>, Fernando Tsukumo<sup>88</sup>, Valéria Valente<sup>89</sup>, Fernando Soares<sup>90</sup>, Eloisa S. Moreira<sup>91</sup>, Diana N. Nunes<sup>92</sup>, Ricardo G. Correa<sup>93</sup>, Heloisa Zalberg<sup>94</sup>, Alex F. Carvalho<sup>95</sup>, Luis F. L. Reis<sup>96</sup>, Ricardo R. Brentan<sup>97</sup>, Andrew J. G. Simpson<sup>98,99</sup>, and Sandro J. de Souza<sup>98</sup>

<sup>1</sup>Ludwig Institute for Cancer Research, 01509-010, São Paulo, Brazil; <sup>2</sup>Departamento de Reumatologia, and <sup>3</sup>Departamento de Biofísica, <sup>4</sup>Faculdade Paulista de Medicina, Universidade Federal de São Paulo (UNIFESP), 04022-062, São Paulo, Brazil; <sup>5</sup>Hemocentro, Universidade Estadual de Campinas, 13089-070, São Paulo, Brazil; <sup>6</sup>Departamento de Radiologia da Faculdade de Medicina da Universidade de São Paulo, 01296-900, São Paulo, Brazil; <sup>7</sup>Departamento de Bioquímica, Instituto de Química, Universidade de São Paulo, 05513-970, São Paulo, Brazil; <sup>8</sup>Departamento de Clínica Médica, Departamento de Biologia Celular e Molecular e Biogênese Patológica, and <sup>9</sup>Departamento de Biogenética e Imunologia, Faculdade de Medicina de Ribeirão Preto, 20001-900, São Paulo, Brazil; <sup>10</sup>Departamento de Genética Biológica, Escola Superior de Agricultura Luiz de Queiroz, Universidade de São Paulo, 13418-878, São Paulo, Brazil; <sup>11</sup>Instituto do Coração (INCor), Faculdade de Medicina, Universidade de São Paulo, 05508-006, São Paulo, Brazil; <sup>12</sup>Laboratório de Nutrição e Doenças Metabólicas, and <sup>13</sup>Departamento de Neurologia, Faculdade de Medicina, Universidade de São Paulo, 01246-900, São Paulo, Brazil; <sup>14</sup>Departamento de Ciências Farmacológicas, Faculdade de Ciências Farmacológicas de Ribeirão Preto, Universidade de São Paulo, 14040-912, São Paulo, Brazil; <sup>15</sup>Departamento de Patologia, Faculdade de Medicina Veterinária e Zootecnia, Universidade de São Paulo, 05508-081, São Paulo, Brazil; <sup>16</sup>Departamento de Genética Médica, Faculdade de Ciências Médicas, Universidade de Campinas, 13081-878, São Paulo, Brazil; <sup>17</sup>Departamento de Histologia Embriologia, Instituto de Ciências Biomédicas, Universidade de São Paulo, 81500-000, São Paulo, Brazil; <sup>18</sup>Departamento de Medicina, Universidade Federal de São Paulo, 04029-122, São Paulo, Brazil; <sup>19</sup>Centro de Biotecnologia, Instituto Butantan, 05502-900, São Paulo, Brazil; <sup>20</sup>Instituto de Pesquisa e Desenvolvimento, Universidade do Vale do Paraíba, 12244, São Paulo, Brazil; <sup>21</sup>Hemocentro, Faculdade de Medicina de Botucatu, Universidade Estadual Paulista, 13816-000, São Paulo, Brazil; <sup>22</sup>Departamento de Genética e Evolução, Instituto de Biologia, and <sup>23</sup>Departamento de Patologia Clínica, Faculdade de Ciências Médicas, Universidade de Campinas, 13082-878, São Paulo, Brazil; <sup>24</sup>Departamento de Genética, Instituto de Biociências, Universidade Estadual Paulista, 18070-081, São Paulo, Brazil; <sup>25</sup>Departamento de Ginecologia e Obstetrícia, Escola Paulista de Medicina, 04311-900, São Paulo, Brazil; <sup>26</sup>Departamento de Biologia, Instituto de Biociências, Letras e Ciências Sociais, Universidade Estadual Paulista, 15054, São Paulo, Brazil; <sup>27</sup>Departamento de Genética Biológica, Faculdade de Ciências Farmacológicas de Araraquara, Universidade Estadual Paulista, 14801-882, São Paulo, Brazil; and <sup>28</sup>Departamento de Anatomia Patológica, Hospital A. C. Camargo, 81589-010, São Paulo, Brazil

Edited by Robert H. Waterston, Washington University School of Medicine, St. Louis, MO, and approved August 2, 2001 (received for review April 11, 2001)

Open reading frame expressed sequences tags (ORESTES) differ from conventional ESTs by providing sequence data from the central protein coding portion of transcripts. We generated a total of 696,745 ORESTES sequences from 24 human tissues and used a subset of the data that correspond to a set of 15,095 full-length mRNAs as a means of assessing the efficiency of the strategy and its potential contribution to the definition of the human transcriptome. We estimate that ORESTES sampled over 80% of all highly and moderately expressed, and between 40% and 50% of rarely expressed, human genes. In our most thoroughly sequenced tissue, the breast, the 130,000 ORESTES generated are derived from transcripts from an estimated 70% of all genes expressed in that tissue, with an equally efficient representation of both highly and poorly expressed genes. In this respect, we find that the capacity of the ORESTES strategy both for gene discovery and shotgun transcript sequence generation significantly exceeds that of conventional ESTs. The distribution of ORESTES is such that many human transcripts are now represented by a scaffold of partial sequences distributed along the length of each gene product. The experimental joining of the scaffold components, by reverse transcription-PCR, represents a direct route to transcript finishing that may represent 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 exploitation of the completed draft human genome sequence. The complexity and variability of human gene structure prevents their direct identification within genome sequence, and supporting data from protein and/or transcript sequence are necessary (1–3). The range of estimates of gene numbers that emanated from the analysis of the draft sequence indicates that we were far from defining a complete catalogue of human genes based on transcript evidence available at that time (4, 5). Thus, there was a pressing need for the generation of further transcript sequence to accelerate the attainment of this goal.

This paper was submitted directly Track1 to the PNAS office.

Abbreviations: EST, expressed sequence tag; ORESTES, ORESTES-RT-PCR, reverse transcription-PCR.

See Commentary on page 11880.

Requests for reprints should be addressed to Sandro J. de Souza: sds@ludwig.com.br.

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "advertisement" in accordance with 18 U.S.C. §1734 solely to indicate this fact.

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\*\*\*

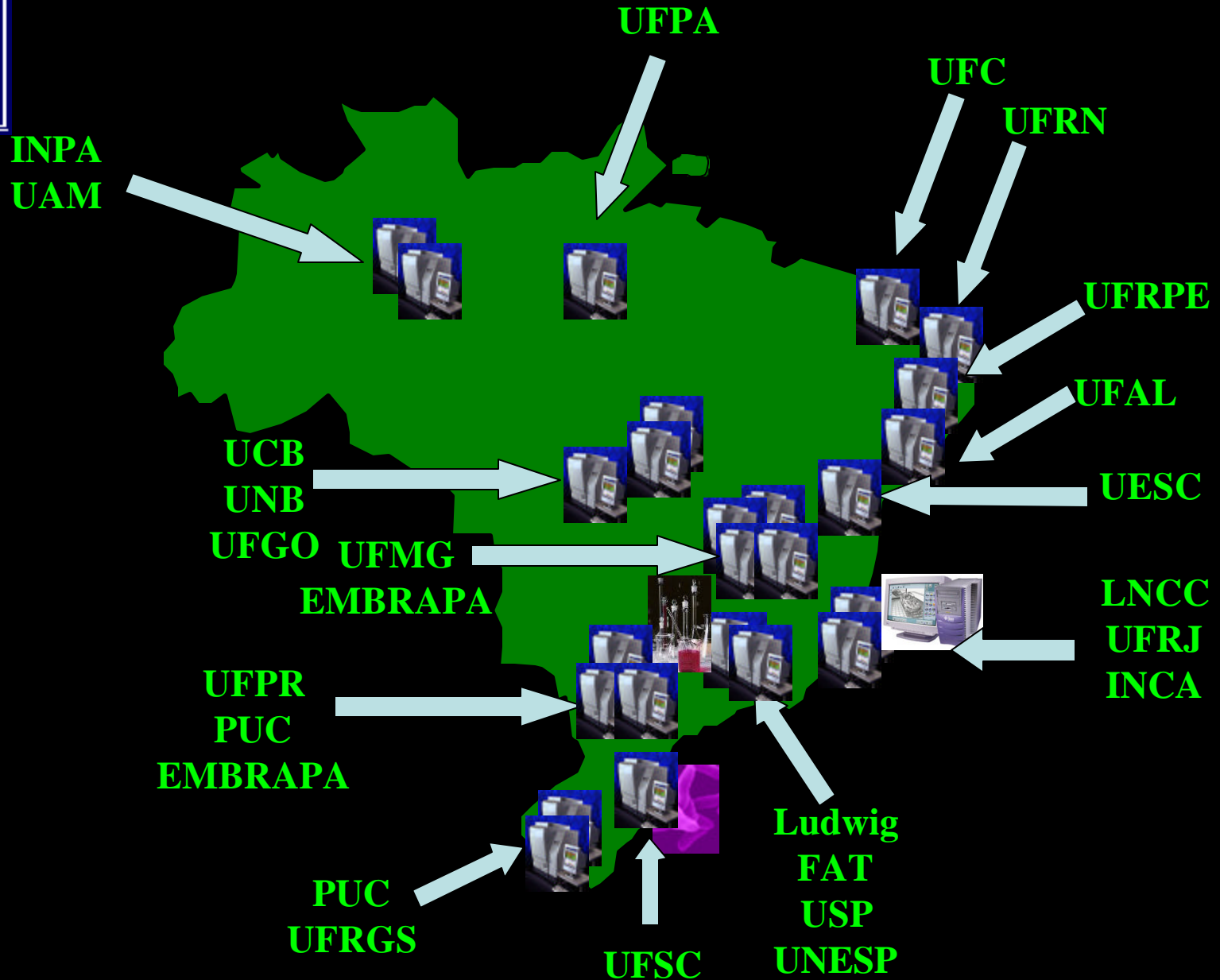
# Organism choice

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



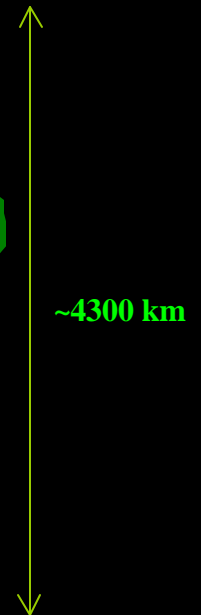
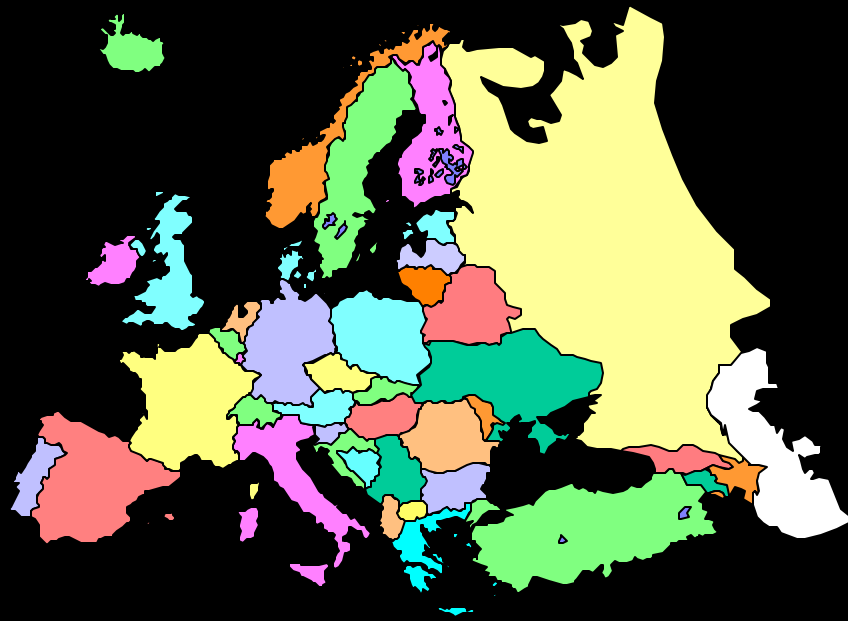
LUDWIG  
INSTITUTE  
FOR  
CANCER  
RESEARCH

# Network



Europe  
9.910.000 km<sup>2</sup>

Brazil  
8.512.000 km<sup>2</sup>



# Regional Genome Projects

## Human Health

*Rede do Nordeste*

*Leishmania chagasi*

*Rede do Centro-oeste*

*Paracoccidioides brasiliensis*

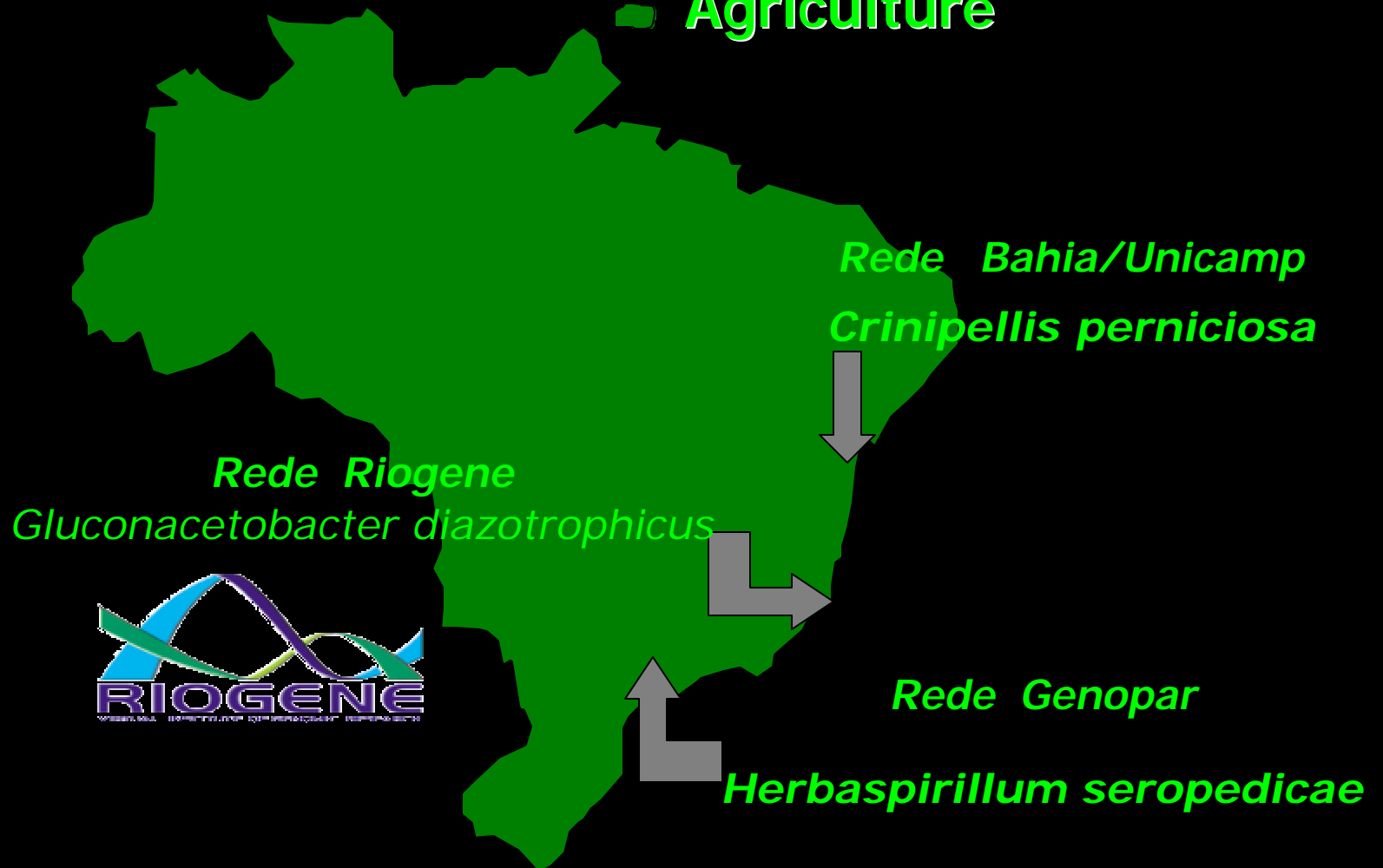
*Rede de Minas Gerais*

*Schistosoma mansoni*



# Regional Genome Projects

■ Agriculture



# 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