Oomycete Genomics, Past, Present and Future

Brett Tyler

Virginia Bioinformatics Institute
Virginia Polytechnic Institute and State University
Systems Biology of Host-Pathogen Interactions

A single genetic network encompasses host and pathogen.
Host-pathogen interactions are influenced by higher scale network connections to...

...microbial communities
...and macroecosystems
Solving the hardest problems in the life sciences requires integration over many scales of organization and multiple domains of knowledge:

- Infection
- Cancer
- Behavior
- Ecology
- Hunger
- Energy
A Wholistic View of Biological Systems

- **20th century genetics**
  - Reductionist
  - One gene-one enzyme
  - Definition of a gene by its knockout phenotype
  - Analysis of small sub-systems one component at a time

- **21st century genetics**
  - Integrative
  - Wholistic view of the genetic potential
  - Large-scale data sets (parts lists)
  - Phenotype as an emergent property of complex systems of component interactions
Oomycete Genome Sequencing

completed or near completed

- *Phytophthora sojae* P6497* (95 Mb) DOE JGI (Sanger)
  - 3 more strains VBI (454)
- *Phytophthora ramorum* (65 Mb) DOE JGI (Sanger)
- *Phytophthora infestans* (240 Mb) Broad Inst (Sanger)
- *Hyaloperonospora arabidopsis* (100 Mb) WUGSC(Sanger), Sanger (Sanger) and Sainsbury (Illumina)
- *Pythium ultimum var. ultimum* (45 Mb) JCVI (Sanger)

*extended contiguity assemblies
Oomycete Genome Sequencing

funded and in progress

• *Phytophthora sojae* P6497 finished sequence* (95 Mb)
  Stanford/HudsonAlpha (Sanger)
  – 3 more strains (P7064; P7074; P7076) VBI (454)
  – 40 million ABI Solid ESTs (ABI/VBI)
• *Phytophthora infestans* supergroup
  – *P. infestans* (2 more strains) Broad Inst. (10x Illumina)
  – *P. phaseoli* U Delaware (Illumina)
  – *P. mirabilis* Broad Inst. (10x Illumina)
  – *P. andina* Broad Inst. (10x Illumina)
  – *P. ipomoeae* Broad Inst. (10x Illumina)
• *Phytophthora capsici* (65 Mb) NCGR/454 (454) and DOE JGI (Sanger)*
• *Pythium ultimum* var. *sporangiferum* (45Mb) JCVI (454)
• *Albugo candida* [10 strains] (40 Mb) Sainsbury (Illumina 50x)
• *Saprolegnia parasitica* (45 Mb) Broad Inst (454, Sanger)
  – *Saprolegnia diclina* (45 Mb) Broad Inst (Illumina 10x)
*extended contiguity assemblies
Oomycete Genome Sequencing

funding being sought

- *Phytophthora parasitica* INRA-310 (Broad host range)* (95 Mb) Broad Inst. (454/Sanger)
  - 11 more strains (Illumina 15-20x)
- *Albugo candida* * (45 Mb) Ag Canada/VBI (454/Sanger/Illumina)
  - assemble reference sequence
- *Aphanomyces euteiches* (legumes)* Genoscope (454)
- *Aphanomyces astaci* * (crayfish)* Genoscope (454)
- *Bremia lactucae* * (45 Mb) (lettuce) UC Davis (454/Sanger/Illumina)

*extended contiguity assemblies*
Oomycete Genome Sequencing

ESTs only

- *Phytophthora parasitica* (3 isolates: BHR, citrus, tomato)
- *Phytophthora brassicae* (*Arabidopsis*)
- *Aphanomyces euteiches* (legumes)
- *Plasmopora viticola* (grape)
- *Lagenidium giganteum* (mosquito larvae)*

* funded, in progress
Oomycete Genome Sequencing Summary

---

**Peronosporomycetidae**

- **Peronospora trifoliorum**
- **Phytophthora ramorum**
- **Phytophthora sojae****
- **Phytophthora capsici**
- **Phytophthora parasitica**
- **Phytophthora infestans***
- **Phytophthora mirabilis**
- **Phytophthora phaseoli**
- **Phytophthora Ipomoeae**
- **Phytophthora andina**

- **Bremia lactucae**
- **Plasmopara viticola**
- **Hyaloperonospora arabidopsidis**

**Pythiales/Lagenidiales**

- **Pythium ultimum****
- **Lagenidium giganteum**

---

**Saprolegniomycetidae**

- **Saprolegnia parasitica**
- **Saprolegnia diclina**
- **Aphanomyces euteiches**
- **Aphanomyces astaci**

---

- high quality assembly
- short read survey
- **** strains sequenced
- genome funding sought
- ESTs
Sequencing of *P. sojae* genotypes

PROGENITORS
- P7074
  - R17
- P1658 & others
  - R1
- P7063, P7064 & others
  - R6 & R7
- P7069, P7076
  - R12 & R19

PROGENY
- P7358
  - R8
- P7082
  - R26
- P7080
  - R24

Group III
- P6955
  - R3
- P3256
  - R3
- P7062
  - R5
- P3258
  - R9
- P7078
  - R21
- P7079
  - R22
- P7303
  - R26
- P7304
  - R36
- P7305
  - R37

Group IV
- P3182
  - R3
- P7061
  - R4
- P7081
  - R25

Group II

CLONES
- P7225
  - P6497
- P1754
- P7359
- P7360
- P7361
- P7364
- P7365
- P7226
  - P7373
  - R4
- P7360
  - R10
- P7361
  - R11
- P7364
  - R13
- P7365
  - R14
- P7470
  - R16
- P7227
  - R2

Group IA

Group I

Oomycete RXLR Effector Proteins

- RXLR-dEER domain enables cell entry independent of the pathogen
- highly variable, under divergent selection
  - 30% of *P. sojae* effectors polymorphic
- 130 to 550 in each sequenced genome
  - *P. sojae* 395
  - *P. ramorum* 350
  - *P. infestans* 550
  - *H. arabidopsidis* 130
  - *Pythium ultimum* ZERO (so far)
- include all 13 cloned oomycete avirulence genes
How RXLR enables effectors to enter cells

Pathogen (fungus or oomycete)

Phosphatidyl-inositol-4-phosphate

Inositol-1,4-diphosphate

Phosphatidyl-inositol-4-phosphate

Host plant cell

Flippase

Endocytosis

Avr1b-GFP alone

Avr1b-GFP+PI-4-P

Avr1b-GFP+Inos1,4-P₂
Data collection in the coming decade

Next generation technologies

- Sequencing, proteomics, metabolomics
- Increased throughput; decreased cost
  - Massively so in the case of sequencing and transcriptomics
- Increased resolution and sensitivity
- Increased integration of stored information
  - Semantic web
  - Drag and drop querying
Next generation sequencing

- **Sanger sequencing with Megabace**
  - $1/1000 bp raw sequence
  - $10/1000 bp 8x assembly
  - 1 machine -> 1 Mb/day
  - 800-900 bp per read + paired end reads

- **454 Life Sciences Pyrosequencing**
  - $0.012/1000 bp raw sequence
  - 1 machine (Titanium) -> 1 Gb/run
  - 400-500 bp per read + paired end reads

- **Illumina/Solexa & ABI SOLID bead sequencing**
  - $0.0005/1000 bp raw sequence
  - 1 machine -> 10-20 Gb/run
  - 50-80 bp per read + paired end reads
Data collection over time and space

Concentrations and locations of nucleic acids, proteins and metabolites

- Sub-cellular localization
- Sub-cellular compartmentalization
- Extra-cellular localization
- Distribution among individuals of a population
- Population genomics and meta-genomics
Sequence analysis will move from snapshots to movies

- Community and population cellular components measured dynamically over space and time
- Comparative and evolutionary analysis at high resolution - no more missing clades
- Model/simulation-driven experimental design to maximize information capture
The Data Challenge

- Collating parts lists
  - sequencing, transcriptomics, proteomics, metabolomics

- Organizing parts lists
  - Databases, cyberinfrastructure, ontologies

- Synthesizing parts lists
  - Pathways, evolution, population genetics
  - Functional genomics
  - Selection of application targets
    - diagnosis, epidemiology
    - disease control targets
Data, data, everywhere but ... 

is it knowledge?

- Even today the ability to generate high throughput sequencing and transcriptomic data is outstripping the ability to transform the data into knowledge.

- Automated data processing pipelines are not a substitute for human insight.
PERL and a Pipetman
Life in a data-rich environment

Every experimental biologist needs to be a computational biologist too

- Theory
- Experiment
- Modeling
- Simulation
Preparing for the Future

Investment and training

- Funding agencies and reviewer communities need to invest more in data analysis.
- Undergraduate and graduate life science programs need to incorporate as much computation as physics.
- Plant pathology programs should exploit natural advantages in this area.
The Intensifying Storm

- New sequencing technologies are exploding and democratizing sequencing capacity
- Microbial genome sequences are rapidly accumulating at sequencing centers, and at dispersed locations
- But - the entire post-sequencing pipeline is being overwhelmed, and value lost, because:
  - the new data are highly heterogeneous, as well as voluminous
  - dispersion creates barriers to integration and synthesis
- Access, integration, and synthesis would be greatly aided by a common, easy-to-use data interface
- Dispersed repositories face special challenges of long-term support and bioinformatics expertise

miGenome.vbi.vt.edu       bmytyler@vt.edu
miGenome

customizable user interface integrating access to diverse repositories and tools

Customizable Community Interfaces

users

admin

WEB SERVICES

Dispersed sequences, tools and services

Existing Sequence Data Repositories

Existing Tool Sets

miGenome.vbi.vt.edu    bmyler@vt.edu
miGenome

customizable user interface integrating access to diverse repositories and tools

Customizable Community Interfaces

admin

users

The information technology to do this already exists. It’s in use in e-commerce and bioinformatics.

Why don’t we already have a common interface?

Existing Sequence Data Repositories

Existing Tool Sets

miGenome.vbi.vt.edu

bmyler@vt.edu
Challenges to a common interface

- Political: ownership and control
  - resource allocation decisions
  - data privacy
- Social:
  - customization of resource
  - It’s more “fun” and easier to build your own
- Professional: career/peer rewards
  - institutions, funding agencies and journals favor individualism and novelty over team play
- Financial:
  - funding agencies, including their reviewer communities, favor individualism and novelty over team play
miGenome Consortium

A community-based initiative

• A common vision: mitigate social, political barriers
  – strong emphasis on the user experience
  – strong emphasis on training new users and user communities
  – easy to plug in new tools and databases
  – unite sequence users, analysts and developers

• A common voice: mitigate financial and professional barriers
  – strengthen advocacy to funding agencies
  – redefine professional expectations

• Advance understanding through synthesis
  – democratize participation
  – addition of value through contributions from diverse communities
miGenome Consortium

Guiding philosophy

“The value of having a highly functional, common genome sequence interface, to the scientific community and ultimately to society in general, supersedes the individual professional interests of any one person or institution”
miGenome Consortium

Current efforts

• Focus initially on fungi and oomycetes
  – greatest unmet need
  – many efforts just starting up
    • US (DOE JGI; several others), UK, Korea; organism db’s
  – political landscape more tractable
  – add protists and algae?

• Identify community members with common concerns

• Work towards a common vision
  – round tables like this one
  – white paper(s)
  – publish open letters

• Involve societies: APS, GSA, ASM

• Talk to agencies: USDA, NSF, DOE, NIAID
  – Microbe Project Interagency Working Group
New paradigm in undergraduate research

Undergraduate consortium in oomycete bioinformatics

VBI, Virginia Tech
Year-round mentoring
Summer Training Workshops

Lafayette College
Manuel Ospina-Giraldo
3 undergraduates

Bowling Green State U
Paul Morris; Karen Sirum
5 undergraduates

College of Wooster, OH
Bill Morgan

Franklin & Marshall College
Jaime Blair

Nova Southeastern College
Aurelian Tartar

funded by NSF/USDA microbial sequencing program
Summary

• 5 oomycete genome sequences are available and several more are on the way
• The rate of new sequence generation is accelerating extraordinarily with next generation technologies
• VBI Microbial Database offers one place to find many sequences
• Rapid accumulation of sequences means new approaches such as miGenome will be needed
• Learning to program greatly enhances your access to diverse sequences
• Undergraduate research in oomycete bioinformatics has much to offer the community, and as a learning tool