Middleware Challenges for the iPlant Collaborative

Dan Stanzione

Co-Director, iPlant Collaborative

Deputy Director, TACC

UT-Austin

iPlant Collaborative™ *Empowering A New Plant Biology*

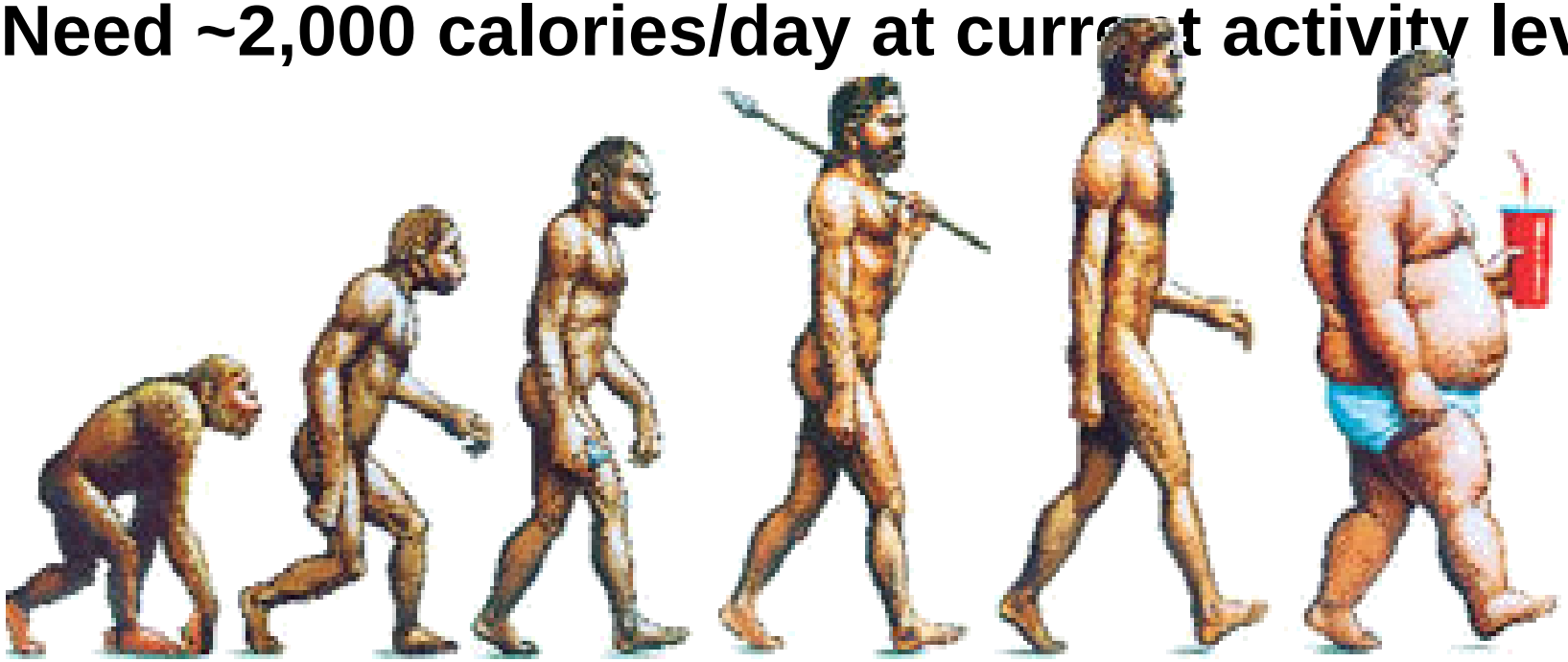
What is iPlant?

- Simply put, the mission of the iPlant Collaborative is to build the CI to support the solution of the grand challenges of plant biology.
- A “unique” aspect is the grand challenges were not defined in advance, but are identified through an ongoing engagement with the community.
- Not a center, but a virtual organization forming grand challenge teams and relying on the nat’l CI.
- Long term focus on sustainable food supply, climate change, biofuels, pharma, etc.
- Now hundreds of participants from around the world… Working group members at more than 50 US academic institutions, USDA, DOE, etc.

iPlant Collaborative™ *Empowering A New Plant Biology*

Changing Lifestyles

- **Activity is being engineered out of life**
- **Adults eat the equivalent of 4 ½ meals a day**
- **US Food production @ 3,800 calories /person/day**
- **Need ~2,000 calories/day at current activity level**



Cereal Consumption

Rice in Asia = 0.9-1.1 lb/day/person

Corn in US = ~3.3 lb/day/person*

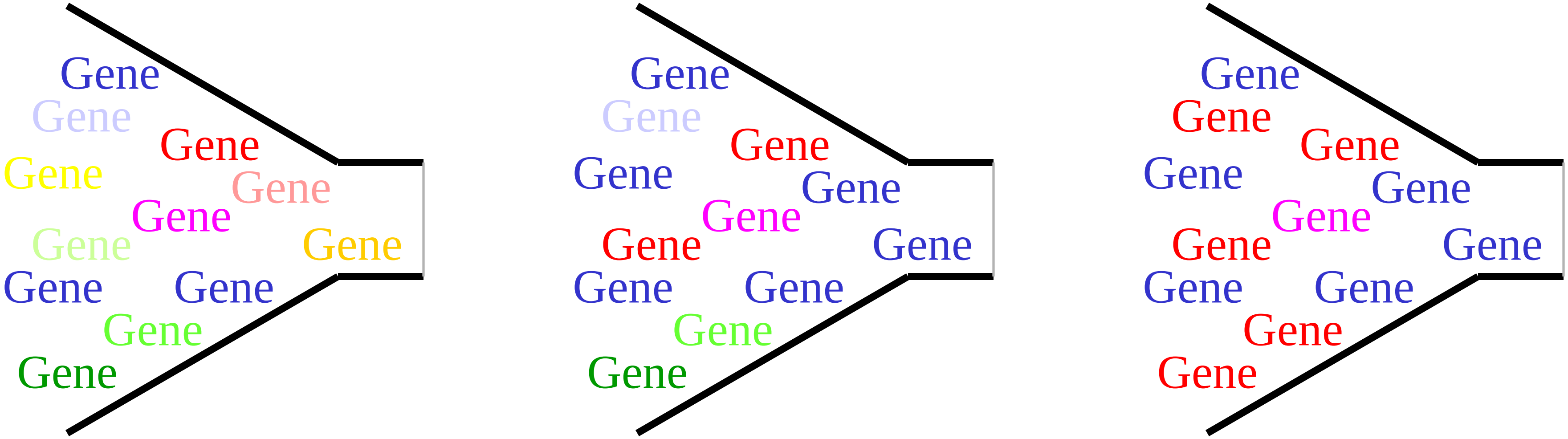
Wheat in Europe = 0.8-1.2 lb/day/person

*** Includes entire food chain**

Wild versus Early Domesticated Corn

Domestication has altered Modern Crops

Domestication Decreases Genetic Diversity



Wild Species → Early Domesticated Species → Modern Varieties

Tanksley and McCouch

Science 1997 August 22; 277: 1063-1066

What is the iPlant CI?

- Two grand challenges defined:
  - iPlant Tree of Life (IPTOL):

    Build a single tree showing the evolutionary relationships of all green plant species on Earth
  - iPlant Genotype-to-Phenotype (IPG2P)

    Construct a methodology whereby an investigator, given the genomic and environmental information about a given individual plant, can predict it's characteristics.
  - Taken together, these challenges are the key to unlocking many "holy grails" of plant biology, such as the creation of drought resistant or pest resistant crops, or breaking reliance on fossil fuel based fertilizer

iPlant Collaborative™ *Empowering A New Plant Biology*

What is the iPlant CI?

- IPTOL CI:
  - Five areas: Data assembly and integration, visualization, scalable algorithms for large trees, trait evolution, tree reconciliation
- IPG2P CI:
  - Five areas: Data Integration, Visualiztion, Modeling, Statistical Inference, Next Gen Sequencing Tools
- In both, a combination of applying compute resources, developing or enhancing new tools, and creating web-based “discovery environments” to integrate tools and facilitate collaboration.

iPlant Collaborative™ *Empowering A New Plant Biology*

What is your single most important cyberinfrastructure challenge/wish?

- Without a doubt, **Data Integration**.
  - Hundreds (at least) disparate data sources
  - Nothing resembling common ontologies

    Not even broad agreement on species names
  - Nothing resembling interoperable data formats
  - No notions of data quality

    This last is perhaps more important in life sciences than any other field I've encountered. Annotations could have been done by anyone

iPlant Collaborative™ *Empowering A New Plant Biology*

Plant Science CI Needs

My personal observations about the plant biology community:

- Data is more important than in other "traditional" CI fields I have dealt with. Truly a "data driven" science.
  - No choice, because there are virtually no models!
  - Simulation is almost a non-issue at this point.
- Corrolary: Communication is harder than in physical sciences, as there is no common mathematical foundation, only a rapidly evolving vocabulary.
  - E.g. "Transposon" (defined in last 3 years).

- *A successful plant science CI will look very different from the type of CI that has had so much impact in other disciplines (e.g. computational chemistry, particle physics).*

iPlant Collaborative™ *Empowering A New Plant Biology*

How do you commission and deploy new CI tools?

- Limited experience in this project with commissioning so far, but here is the *plan*:
  - For in-house tools, rigorous testing incorporated into development from square one; we have full time testers who run them through system test before deployment.
  - For 3rd party tools:
    - Limited testing for system stability before acceptance
    - Initial release by users makes them available but in “untested” category.
    - More rigorous testing before receiving “blessed” status, including evaluation of accuracy, and community feedback
    - Multiple recommendations to rank tools via a community rating system; we will likely do this.

iPlant Collaborative™ *Empowering A New Plant Biology*

What has been so far your biggest CI failure/disappointment if any?

- A little early to answer this as “real” development basically started in August ‘09
- It has been *extremely* difficult to get the community to even approach agreement on what the CI needs are, even within fairly focused groups who agree on the grand challenge questions.
- Fuzzy notions of what CI is don’t help.