# iPToL_year4_roadmap

## Data Assembly

### General Data Assembly

**Deliverables:**

- An industrial strength pipeline for data assembly for very large data sets;
- Collaboration and Analysis Tools (contribute one's own data, download data, analyze data, keep data/results private) through the DE.

**Strategy**:

- Engage iPlant faculty (Nirav Merchant, Sudha Ram, Eric Lyons) for domain expertise in data infrastructure, meta-data management and scientific workflows.

**Tasks**:

- Robust data upload capability (IRODS; Rion Dooley, Nirav Merchant), late 11Q1?
- Meta-data management (Rion Dooley, Nirav Merchant, Sudha Ram), early 11Q2;
- Robust data storage and retrieval, collaboration tools in DE (iPlant-wide requirement; core software), 11Q2;
- Advanced collaboration tools (iPlant-wide requirement; core software), 11Q3;
- Input data validation (core software in collaboration with data integration ETAs);
- Multiple sequence alignment generation:
    - PHLAWD (John Cazes, Stephen Smith);
    - Gordon Burleigh's pipeline (John Cazes, Eric Lyons, Stephen Smith);
    - Muscle, other alignment strategies (John Cazes, Eric Lyons).
- Sequence database(s) bad GIs for PHLAWD, etc (Sheldon McKay and delegates), 11Q2.

### My-Plant/My-Crop

**Deliverables:**

- My-Plant: Robust, widely used scientific collaboration network for the plant sciences based on a phylogeny metaphor;
- My-Crop: In support of integrated breeding platform, a scientific interaction site as well as data landing pad.

**Strategy**:

- Build a phylogenetically-structured social networking website for information sharing and collaboration;
- Generalize and extend to other display/networking paradigms.

**Milestones**:

1. Official launch 10Q3.

**Status**:

- 239 users, 57 clades (Feb 3, 2010).

**Tasks:**

- Refactor back-end for generic 'clade' structure to facilitate other display/organization paradigms (Matther Helmke), late 11Q1:
    - Node based;
    - Drupal core taxonomy.
- Implement Drupal module to ingest and provide basic search functionality for relevant literature citations from the user community and public databases (Steve Mock and delegates), 11Q2;
- Integration with Facebook and other similar sites. Initially with passive linkouts, possible later with Facebook page or app (Steve Mock and delegates), 11Q3;
- Integration (as consumer) of TNRS and iPlant tree viewer (Steve Mock; Matt Hanlon);
- My-Crop (different paradigm for display – also data repository):
    - scoping, early 11Q2;
    - early implementation, late 11Q2;
    - full implementation, early 11Q3.

## Trait Evolution

**Deliverables**:

- An infrastructure for trait analysis and ancestral characters estimation.

**Strategy**:

- Integration of limited number of trait analysis tools with special focus on R scripts (since Sept 2010).

**Milestones**:

1. Identified set of 5 components for integration:

- Phylogenetically Independent Contrast (PIC);
- Discrete Ancestral Character Estimation (DACE);
- Continuous Ancestral Character Estimation (CACE);
- Tree stretching models;
- Discrete traits correlations (Pagel 94).

2. Released 1st component (PIC) 10Q2.

**Status**:

- DACE and CACE included in the 3rd release of DE 11Q1;
- Code improvements in 2 existing R programs (ape, geiger), will be pushed back to community, 11Q1;
- Tree stretching and Pagel94 to be included in 4th DE release 11Q2.

**Tasks:**

- Tool integration in DE-R scripts and command line tools (Naim Matasci and delegates), early 11Q2;
- Data uptake- files and external web based data from TreeBase (Sonya Lowry and delegates);
- Tree viz integration:
    - New visualization needs (Kris Urie), early 11Q2;
    - Call backs for new analyses (Adam Kubach, Sonya Lowry), mid 11Q2.
- DE integration:
    - Integration (Sonya Lowry and delegates), late 11Q2;
    - Metadata mapping (Naim Matasci), late 11Q2;
    - Analysis and viewer integration (Sonya Lowry and delegates), late 11Q2.
- Code and documentation release (Naim Matasci, Matthew Helmke), 11Q2.

# Tree Reconciliation and onekp

**Deliverables**:

- Applications to perform, visualize and analyze the evolution of gene families from the onekp project with gene-species tree reconciliations.

**Strategy**:

- Development of an analytical pipeline, a database schema and a visualization tool;
- Populate with data from onekp project.

**Milestones**:

1.Bioinformatic pipeline for gene-species tree reconciliation completed and database populated with the reconciled trees, 10Q4;

2.Developer preview released with the following features, 11Q1:

- Database containing reconciliations for over 2500 gene families in six examplar species (poplar, grape, cucumber, papaya, soybean and Arabidopsis thaliana), 11Q1;
- GUI with the ability to search and view reconciled trees and to download data;
- Display of species trees and gene trees side-to-side, using the Tree Visualizer developed by the Tree Viz Working Group;
- Interactive mapping of duplication and speciation events between gene and species tree and vice versa;
- Markups for speciation and duplication events on the gene tree nodes and of duplication events on the species tree branches;
- Ability to add additional markups;
- Contextual menus;
- Advanced search functionality, including:
    - BLAST,
    - GO terms and IDs,
    - Gene IDs;
- GO tag clouds for gene families;
- Retrieval of underlying data (sequences and reconciliations).

**Status**:

- Ready to receive onekp data 11Q1;
- All the current onekp assemblies (36,998,590 assemblies from 398 total species/tissues) mirrored at TACC;
- Have run blastx searches on each of the onekp assemblies (complete);
- Blast server at TACC for onekp consortium members to search against the assemblies using either single sequences or batches (complete).

**TR Tasks:**

- Data modeling and database schema design (Sheldon McKay, Jamie Estille), 10Q3, update 11Q2;
- Port tree reconciliation analysis pipeline to TACC HPC resources (Sheldon McKay), early 11Q2;
- Adapt tree reconciliation analysis pipeline to use Bayesian tree building method (PrIME-GSR, Sheldon McKay), mid 11Q2;
- Update stand-alone web application for onekp/TR data (Naim Matasci), early 11Q2;
- DE integration (Sonya Lowry), 11Q2;
- Establish/negotiate onekp data release policy (Sheldon McKay,Jim Leebens-Mack and Gane Wong), 11Q1;
- Expose services through DE and API (TBD), 11Q2;
- Release code and documentation (Matthew Helmke), 11Q1;
- Adjustments to viz as required (Adam Kubach, Kris Urie).

**Onekp Tasks:**

- Sequencing and transcript assemblies (onekp consortium; external);
- Gene cluster identification and alignment (Norm Wickett), ongoing;
- Continue to manage onekp data intake and tool implementation at TACC (Michael Gonzales, Sheldon McKay, Chris Jordan).

# TNRS

**Deliverables**:

- A Taxonomic Name Resolution Service that:
    - will query taxonomic data from Tropics and other data services using GNI architecture and global names index allowing for different nomenclatures;
    - recover validated names names using exact and fuzzy matching algorithms;
    - inspect taxonomic status of validated names and convert synonyms where applicable.

**Strategy**:

- Development of a tool based on TaxaMatch (Tony Rees, CSIRO Marine and Atmospheric Research) and GNI Parser (Dmitry Mozzherin from the Encyclopedia of Life).

**Milestones**:

1. First release of the tool 10Q4;
2. Completion of Phase 1/Scoping of Phase 2 11Q1;
3. Support for Family and infraspecific epithets 11Q1.

**Status**:

- Active development towards synonymy support and ability to use other data sources, 11Q2;
- Improvements to GNI parser and TaxaMatch codes pushed back to community, 11Q1.

**Tasks**:

- UI redesign (Nicole Hopkins), 11Q1;
- Algorithm to handle synonyms (Jerry Lu), early 11Q2.

# Big Trees

**Deliverables:**

- Computational infrastructure to build ToL.

## NINJA/WINDJAMMER.

**Strategy**:

- Optimization of NINJA (neighbor joining implementation) for HPC.

**Milestones**:

1. Software rewritten from Java to C with an MPI;

2. On-board distance matrix calculation added (K2P and Jukes Cantor for DNA; Blossum 42 for protein);
3. Six day run time reduced 32-fold to 4.5 hours for 220K species data set;
4. Two/three day run time reduced 1,800-fold to 2 minutes for distance matrix calculation on 220K set.

**Status**:

- Completed, minor tweaking of MPI.

## RAxML

**Tasks:**

- Implement RAxMl-lite on Ranger, benchmark with various data sets (John Cazes), 11Q1;
- Implement web interface (relies on foundational API; Steve Mock and delegates), 11Q2.

**Status**:

- In progress

## Phylogenetics Workflow and Perpetually Updating Tree

**Workflow**

A Nascent workflow has been added to the the DNA subway as an education tool. This can serve as a model for integrating phylogenetic analysis tools to the DE.

**Perpetually updated TOL**
This is predicated on the completion of the infrastructure for data matrix assembly, RAxML-lite tree building, tree vizualization etc and is being scoped by the the iPlant scientific project Management team (Eric Lyons, Sheldon McKay, Matt Vaughn, Nicole Hopkins). Advice will be sought from the iPToL faculty regrading further requirements.

- The basic strategy is an automated workflow that will synch with GenBank or other data repository, build or iterate on on a chaacter matrix, re-run the tree building and update the Discovery Environment.

# Tree Visualization

**Deliverables**:

- An interactive tree viewer that:
    - Makes possible to view large trees as a stand alone tool;
    - Makes the green plant ToL and sub-trees available in the iPlant DE;
    - Meets the visualization needs of Trait Evolution and Tree Reconciliation and of other applications in the Discovery Environment.

**Strategy**:

- Development of a *de novo* tree viewer using GWT.

**Milestones**:

- Tree able to display 500K taxa with semantic zooming;
- Search capabilities;
- Metadata driven node interactions;
- Visual annotations (node and branch colors, thickness, size, etc.);
- Designed API for generalized use;
- User interactions support;
- User interface.

**Status**:

- Completed development for TR functionality and moved into maintenance mode (pending changes needed to accommodate large datasets). To be included with one release (late 11Q1);
- Live demoed to Trait Evolution Working Group.

**Tasks:**

- Continuing development towards inclusion of TE functionality in the 4th DE release (Adam Kubach, Kris Urie), 11Q2;
- Integration into the DE (Sonya Lowry), 11Q2;
- Standalone release (Karen Cranston), 11Q3;
- Code and documentation release (Karen Cranston, Matthew Helmke), 11Q2.