The Maize Rare Alleles Project: Biology & Bioinformatics

Jeff Glaubitz

Senior Research Associate, Buckler Lab,
Institute for Genomic Diversity, Cornell University

Project Manager of Panzea (The NSF Maize Diversity Project)

Bioinformatics Practitioners Club
April 20, 2015

The Maize Rare Alleles Project: Biology & Bioinformatics

1. **Biological Goals of the Project**
   - Better prediction of the effects of rare alleles
   - Biology-assisted breeding
   - Accerated breeding of maize & other crops

2. **TASSEL Overview**
   - User resources
   - Developer resources

3. **TASSEL-GBS Pipeline**
   - v1 and v2

4. **Maize Genome Annotations DB**
   - Initial attempt and ongoing challenges

Biology of rare alleles in maize and its wild relatives

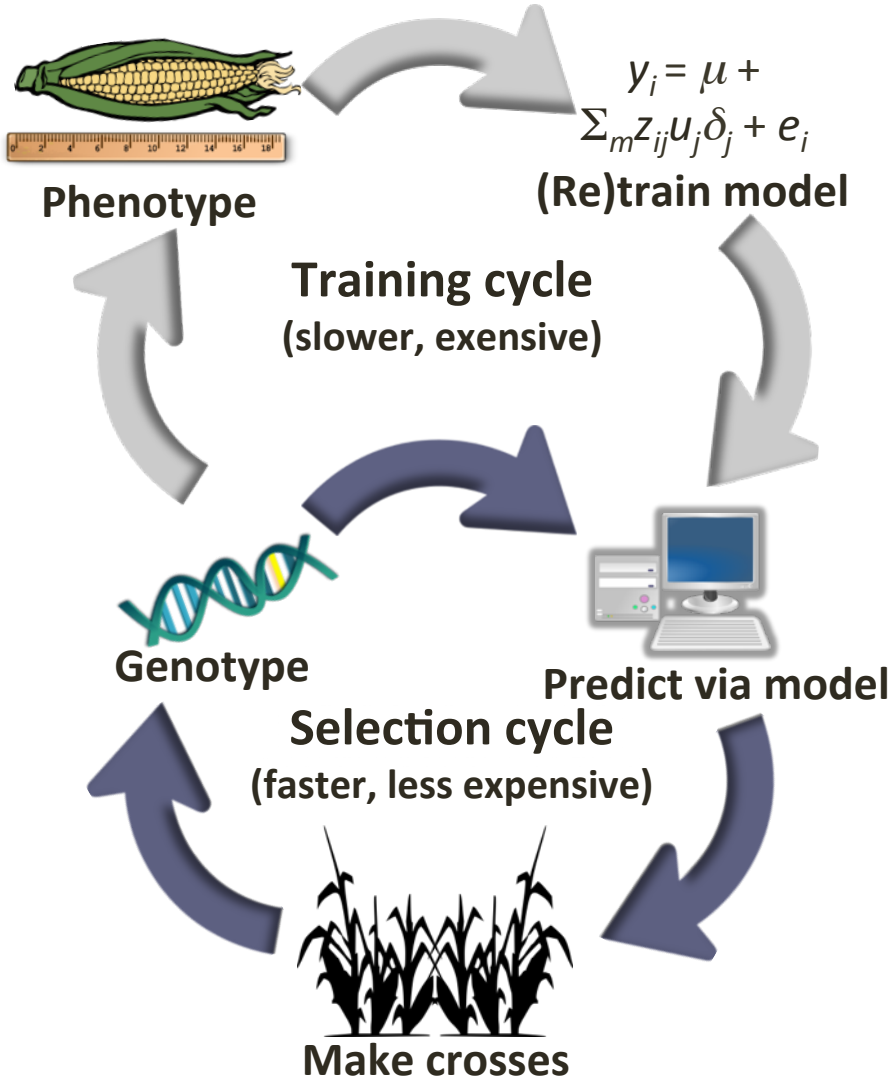
Cornell:

- **Edward Buckler** (**PI**, USDA-ARS/Cornell)
- **Peter Bradbury** (USDA-ARS, Ithaca, Analysis)
- **Qi Sun** (Bioinformatics)
- **Sharon Mitchell** (GBS Genotyping)
- **Theresa Fulton** (Outreach Coordinator)
- **Jeff Glaubitz** (Project Manager)

- **John Doebley** (University of Wisconsin)
- **Sherry Flint-Garcia** (USDA-ARS, University of Missouri)
- **James Holland** (USDA-ARS, North Carolina State University)
- **Jeffrey Ross-Ibarra** (University of California, Davis)
- **Doreen Ware** (USDA-ARS, Cold Spring Harbor Lab)

- 15 postdocs, 13 graduate students, numerous undergrads

Genomics Assisted Breeding



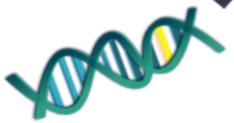
Biology Assisted Breeding

Phenotype

$$y_i = \mu + \Sigma_m z_{ij} u_j \delta_j + e_i$$

(Re)train model

Training cycle
(slower, exensive)

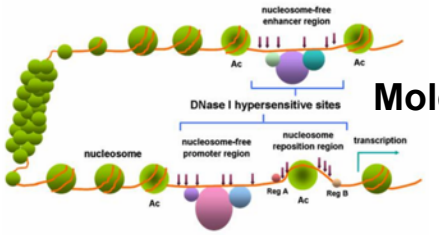


Genotype

Predict via model

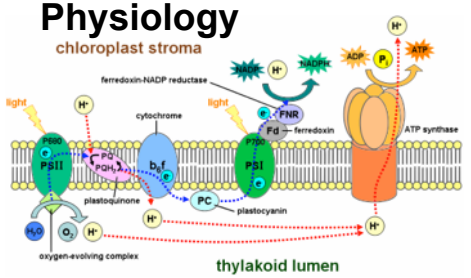
Selection cycle
(faster, less expensive)

Make crosses

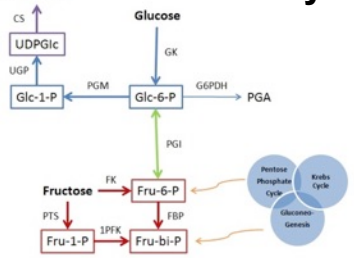


Molecular biology

Physiology



Biochemistry



Evolution

Environment

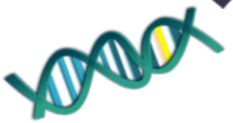
Biology Assisted Breeding

Phenotype

$$y_i = \mu + \Sigma_m z_{ij} u_j \delta_j + e_i$$

(Re)train model

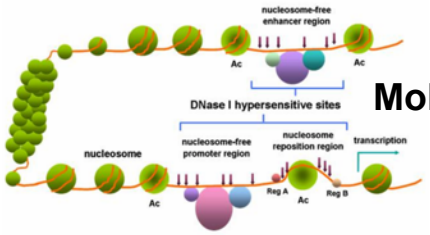
Training cycle
(slower, exensive)



Genotype

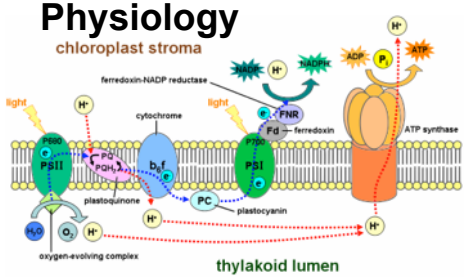
Predict via model

Selection cycle
(faster, less expensive)

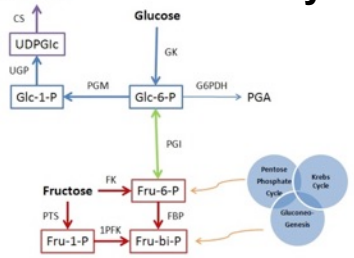


Molecular biology

Physiology



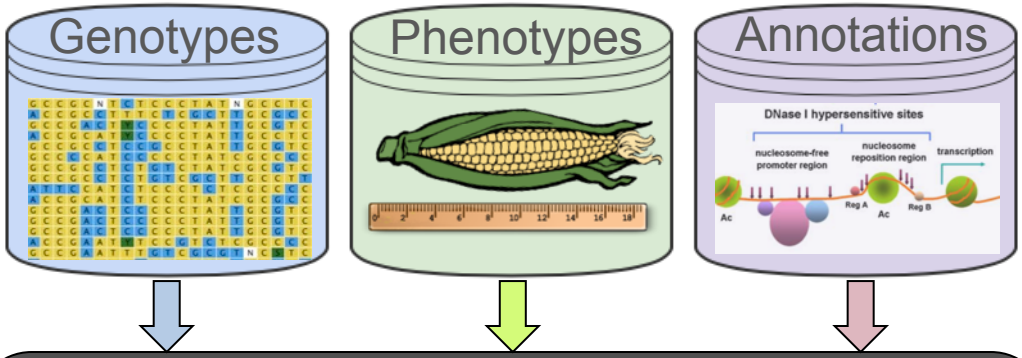
Biochemistry



Evolution

Environment

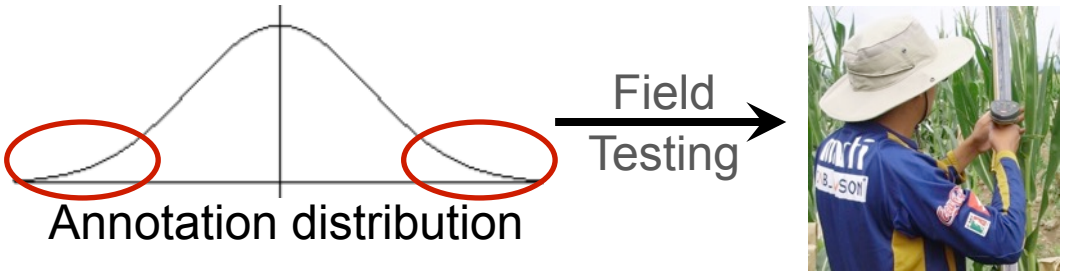
Data:



Analysis:

GWAS
Machine Learning
Variance (h^2) Explained
Enhanced Genomic Prediction

Verification:

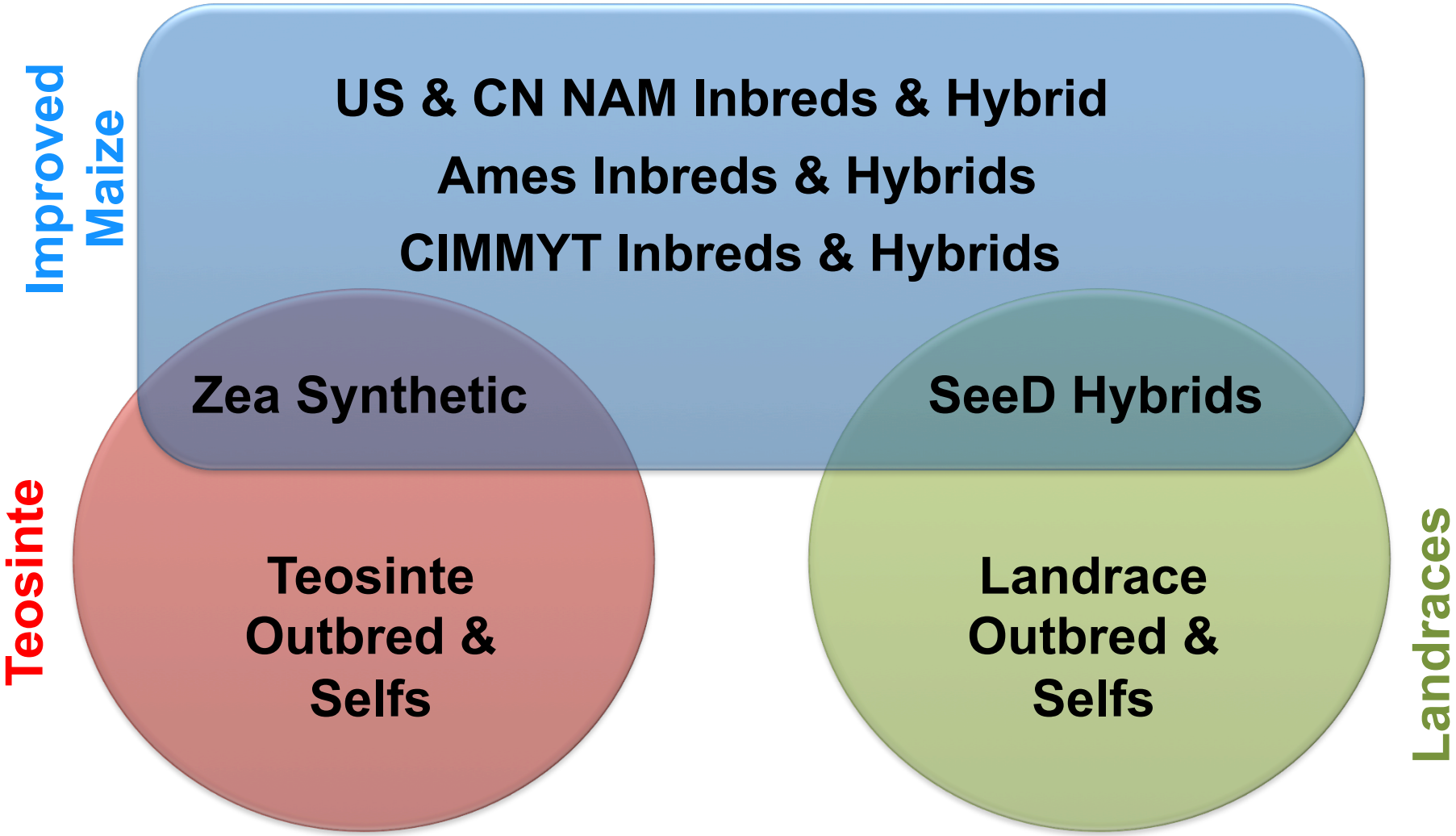


Products:

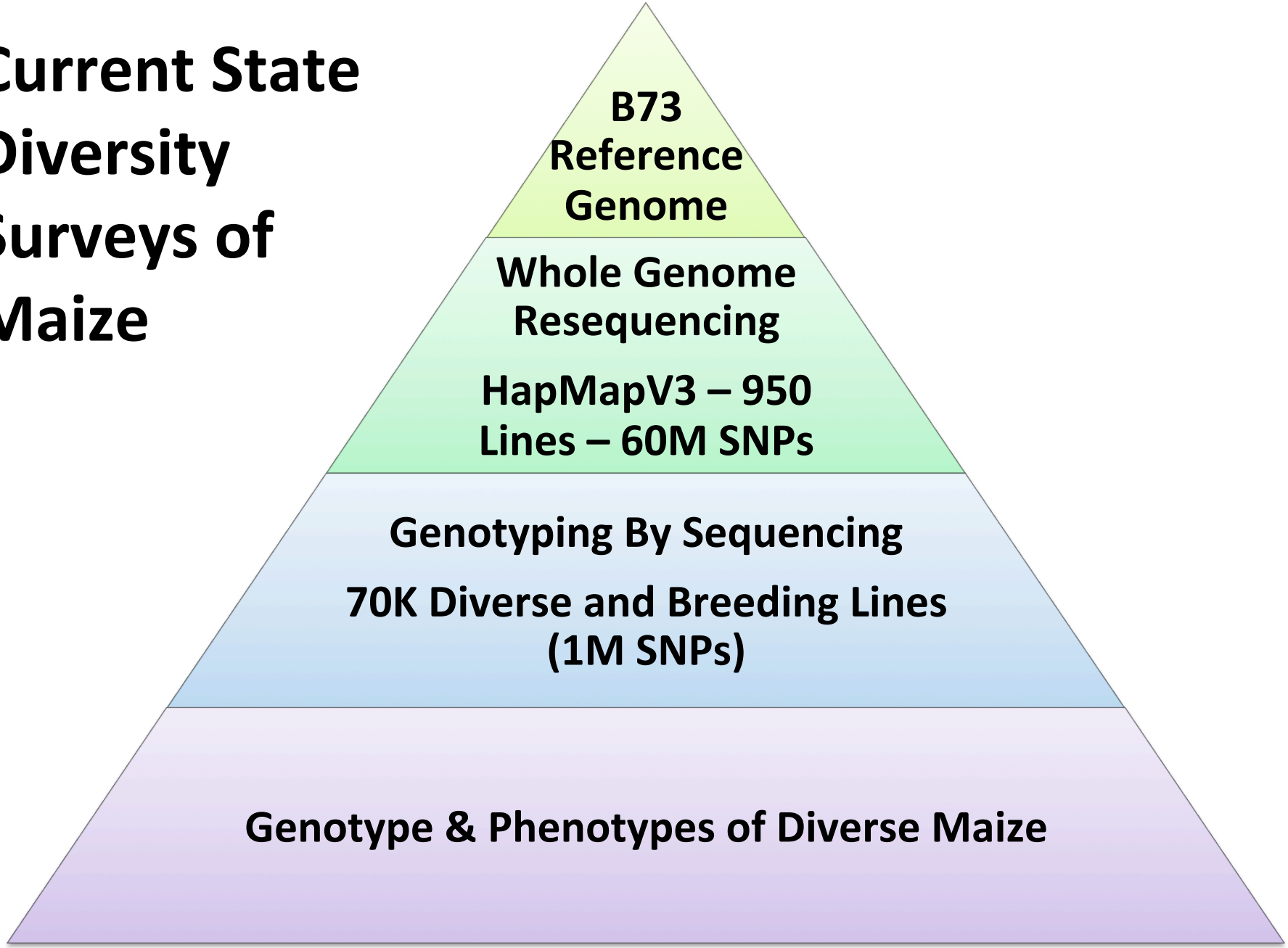
Accelerated Breeding

Personalized Medicine for Corn?

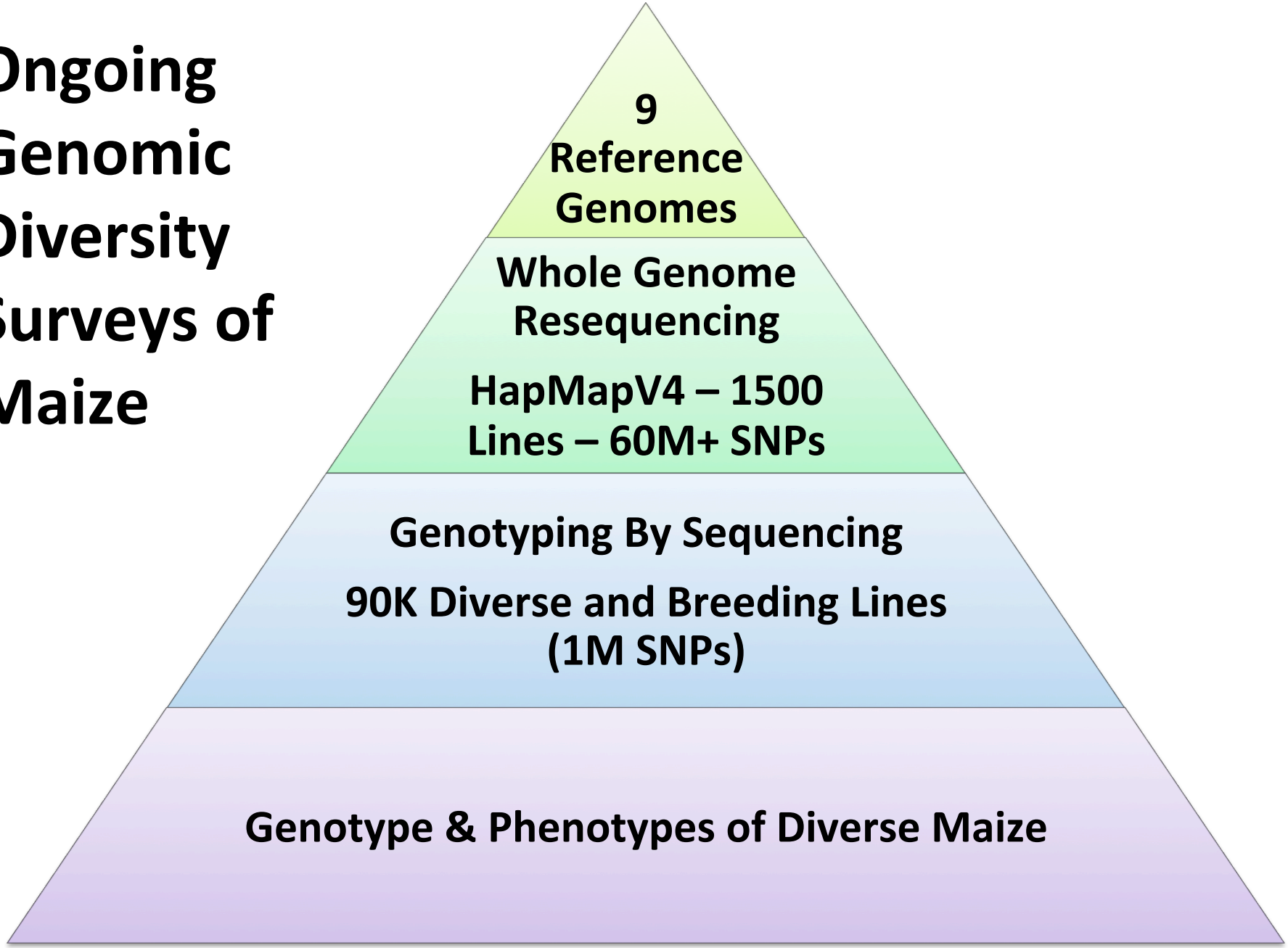
The Genomic & Field Data Sample a Range of Historical Inbreeding with a Variety of Dominance Relationships



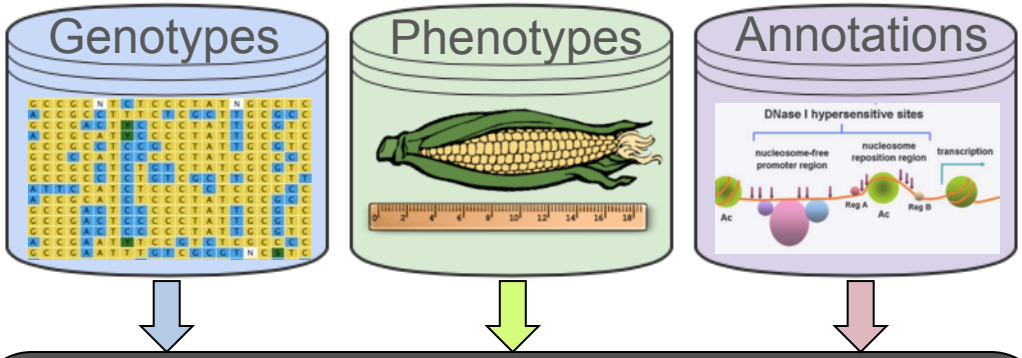
Current State Diversity Surveys of Maize



Ongoing Genomic Diversity Surveys of Maize



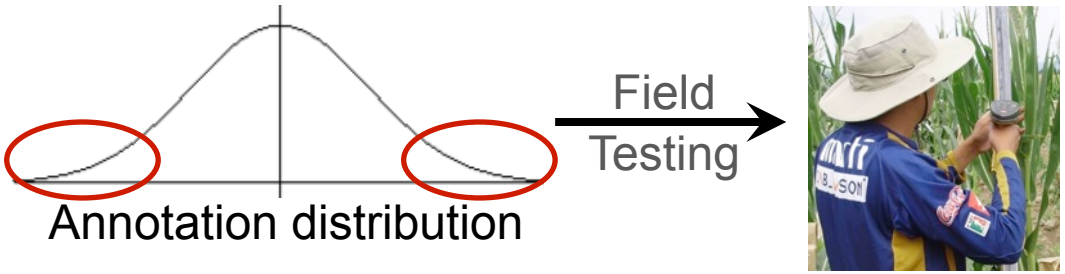
Data:



Analysis:

GWAS
Machine Learning
Variance (h^2) Explained
Enhanced Genomic Prediction

Verification:



Products:

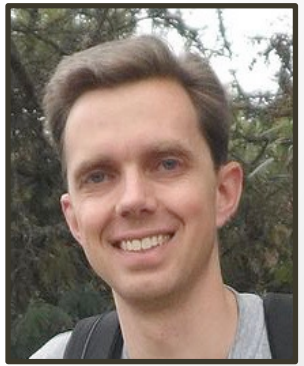
Accelerated Breeding

Personalized Medicine for Corn?

What is TASSEL?*

- TASSEL's primary purpose is to help the Buckler lab do what the Buckler lab wants to do
- Not originally intended as a community resource
  - Contrast with Bowtie, SOAP, etc.
- Basically a collection of useful tools rather than a unified framework

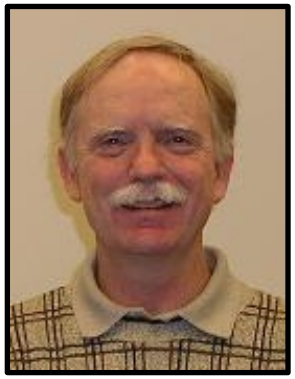
*Slides with this format courtesy of Jason Wallace:



Authors of Tassel

Ed Buckler

Terry Casstevens



Peter Bradbury

Lynn Johnson

Zak Miller

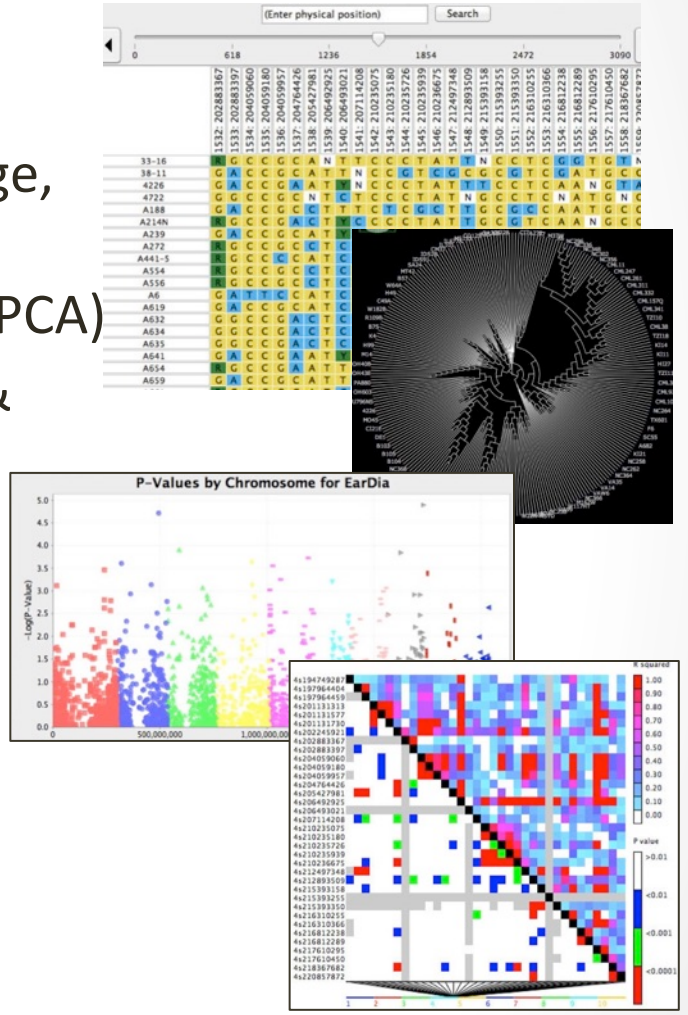
Kelly Swarts

Fei Lu

Jeff Glaubitz

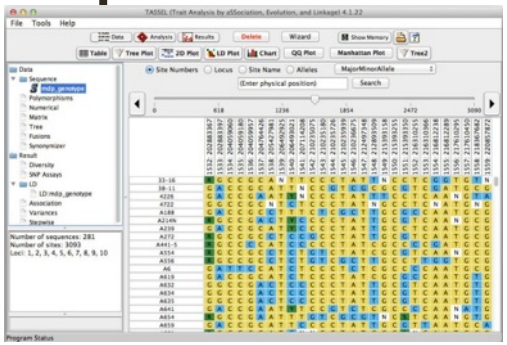
What TASSEL Can & Can't Do

- What TASSEL can do:
  - Manipulate genotypes (filter, merge, etc)
  - Basic popgen (phylogenetic trees, PCA)
  - Association analysis (GLM, MLM, & components)
  - GBS SNP calling
  - Imputation (FILLIN or FSFHap)
- What TASSEL can't do:
  - Other imputation algorithms
  - Advanced popgen
  - Normal linkage mapping
  - And lots of other things



Ways to work with TASSEL

Graphical Interface



Command Line

```
> run_pipeline.pl $TASSEL -fork1 -h
allzea_gbs_2_7.t5.h5 -filterAlign -
filterAlignMinFreq -includeTaxaInFile
my_target_taxa.txt -export
mytaxa_gbs_2_7_filtered.hmp.h5 -
runfork1
```

API (Java)

```
Public class FilterMyStuff{

    GenotypeTable myGenos =
        ImportUtils.ReadFromHapmap("
        test_gbs.hmp.txt.gz")
    FilterGenotypeTable filtered =
        FilterGenotypeTable.getInstance
        (myGenos, myTaxaList);
    …
```

Intuitiveness

Power

File Formats

- **HapMap** (.hmp.txt)
  - Relatively compact 2-way table; only stores genotypes
  - TASSEL's default format
- **VCF** (Variant Call Format; .vcf)
  - Much more data rich (and bigger files)
  - Can store many pieces of data at each position
  - TASSEL can read/write, but often lossy
- **HDF5** (Hierarchical Data Format 5; .h5)
  - Generic format for accessing large datasets
  - Basically a mini file system within a file
  - *Not* human-readable
  - TASSEL's (newly) secondary format

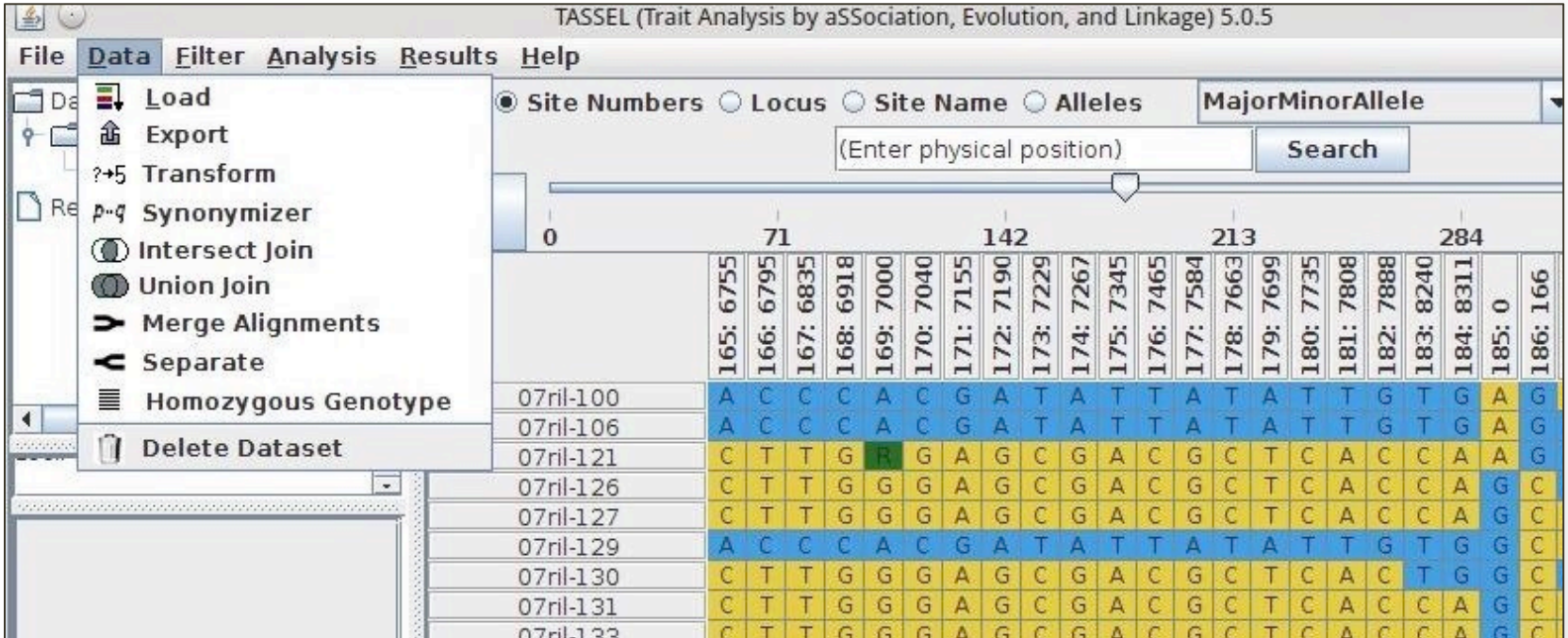
NCSA BOEING

NOAA NATIONAL OCEANIC AND ATMOSPHERIC ADMINISTRATION U.S. DEPARTMENT OF COMMERCE

NASA

General Philosophy - GUI

- Organize everything by menus
- Fairly intuitive
  - Only tricky bit is that loading files happens under “Data”, not “File”



General Philosophy – Command Line

- Command line *much* more powerful than graphical interface
- Basic idea is to chunk commands into groups and pass data from one group into another
- Can do independent tasks in parallel

Command-line organization

```
> run_pipeline.pl -Xms2g -Xmx8g -fork1 -vcf
   myfile.vcf -includeTaxaInFile 0_core_taxa.txt
   -export myfile_filtered.hmp.txt.gz -runfork1
```

Command Line Organization

```
> run_pipeline.pl
  -Xms2g -Xmx8g
  -fork1
  -vcf myfile.vcf
  -includeTaxaInFile 0_core_taxa.txt
  -export myfile_filtered.hmp.txt.gz
  -runfork1
```

Perl script to start TASSEL

Java arguments to set memory

Start of first command set

Load file

Filter data

Export result

Run first command set

Plugins
The hidden underbelly of TASSEL

- TASSEL is filled with “plugins” – mini-programs to do various things
  - E.g., the GBS pipeline is a series of related plugins
- Invoked with following code:

```
-forkX –NameOfPlugin [arguments] –endPlugin -runforkX
```

- If no arguments given, it *should* print out a list of expected ones
- Currently no way to see a list of all available plugins

XML exporting

- TASSEL includes utility to save an XML record of your command, which can be used again directly
- Not automatable, but still useful
- **Create XML file**
  - `run_pipeline.pl -createXML config.xml -fork1 ...`
- **Run from XML file**
  - `run_pipeline.pl -configFile config.xml`
- **Translate XML back to command line (doesn't run)**
  - `run_pipeline.pl -translateXML config.xml`

Scripting

- Script = test file that a program can interpret to run something
- Good for automating TASSEL commands
  - Also, keeps a record of what you ran
- Usually in Bash (Linux/Unix/Mac), but can be others as well

```
#! /bin/bash

TASSEL="/home/user/tassel5-standalone/run_pipeline.pl –Xms2g –Xmx6g"
for pop in Collaborations/pop_*_taxa.txt; do
    $TASSEL –fork1 -h5 maizepalooza.t5.h5 –includeTaxaInFile $pop –filterAlign …
done
```

Resources for Tassel Users

- Available via www.maizegenetics.net/tassel:

TASSEL 5.0

Tassel Version 5.0 *(Getting Started!)*
(Build: April 9, 2015 Requires: Java 1.8)

Tassel 5 Mac OS
Tassel 5 Windows 64 Bit
Tassel 5 Windows 32 Bit
Tassel 5 UNIX

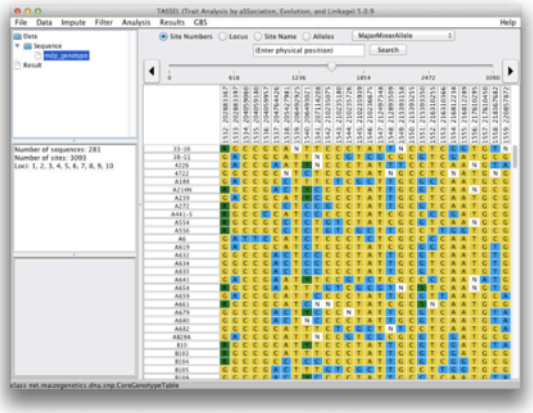
Tassel Version 5.0 Standalone
(GBS Pipeline - Under Development)

Using Git - *Recommended!*
Download *(Click on "Tags")*

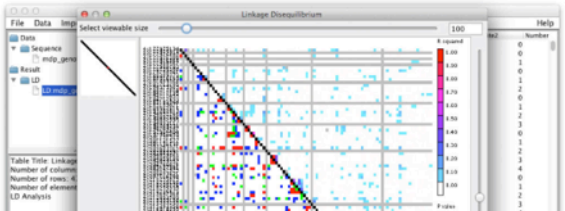
Archived Versions of Tassel
(GBS Pipeline - Stable)

Tassel Documentation

Alignment Viewer



Linkage Disequilibrium Display



Resources for Tassel Users

- Available via www.maizegenetics.net/tassel

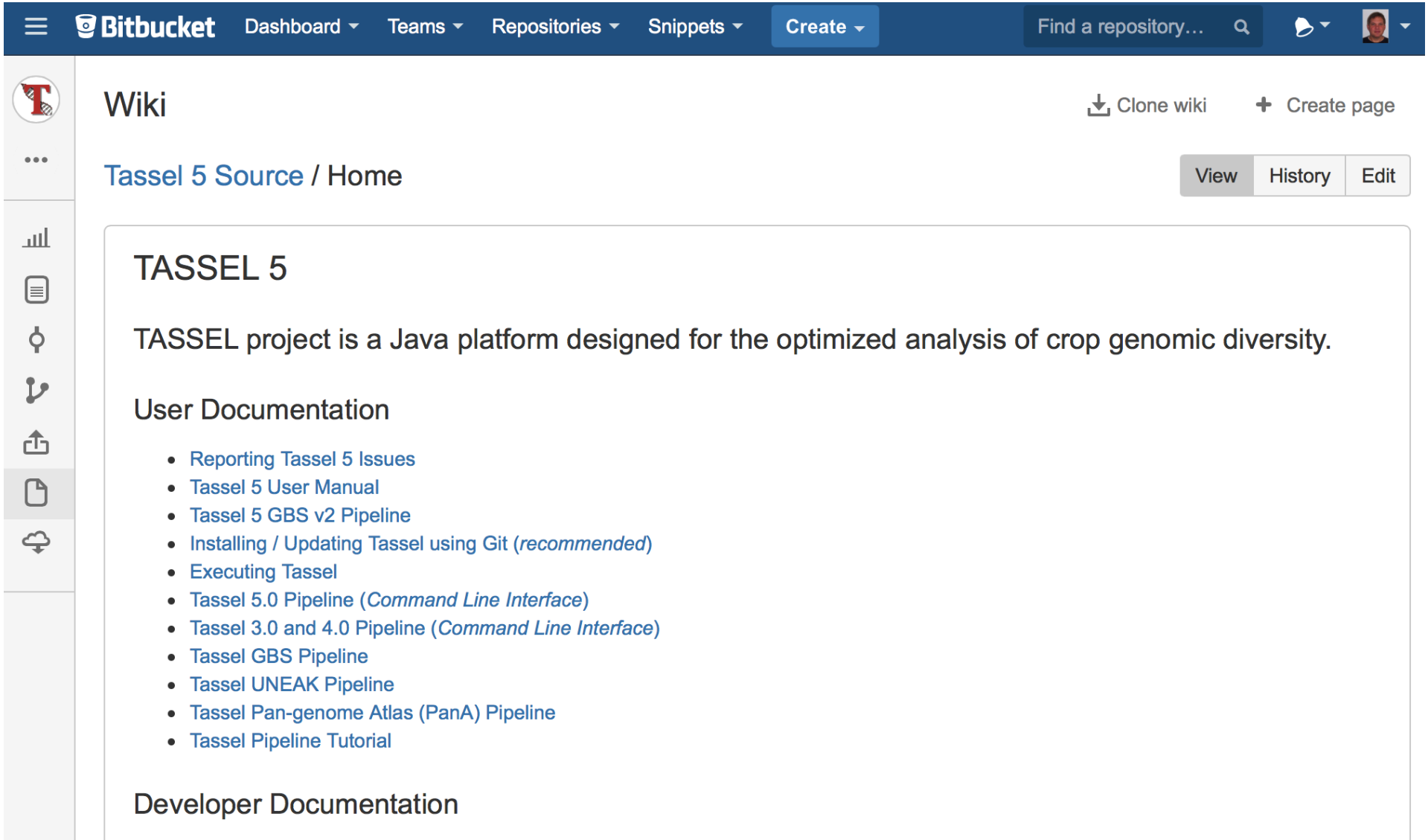
Tassel 5 User Manual

Introduction

- User
  - Web Site
  - Contributors
  - Citations
  - Getting Started
  - Executing TASSEL
  - Open Source Code
  - Software Development Tools
  - Graphical Interface
  - Pipeline (Command Line Interface) *Tassel Pipeline Tutorial*
  - GBS Pipeline
  - UNEAK Pipeline
  - Tassel Pan-genome Atlas (PanA) Pipeline

Resources for Tassel Users

- Available via www.maizegenetics.net/tassel



Resources for Tassel Users

- Available via www.maizegenetics.net/tassel

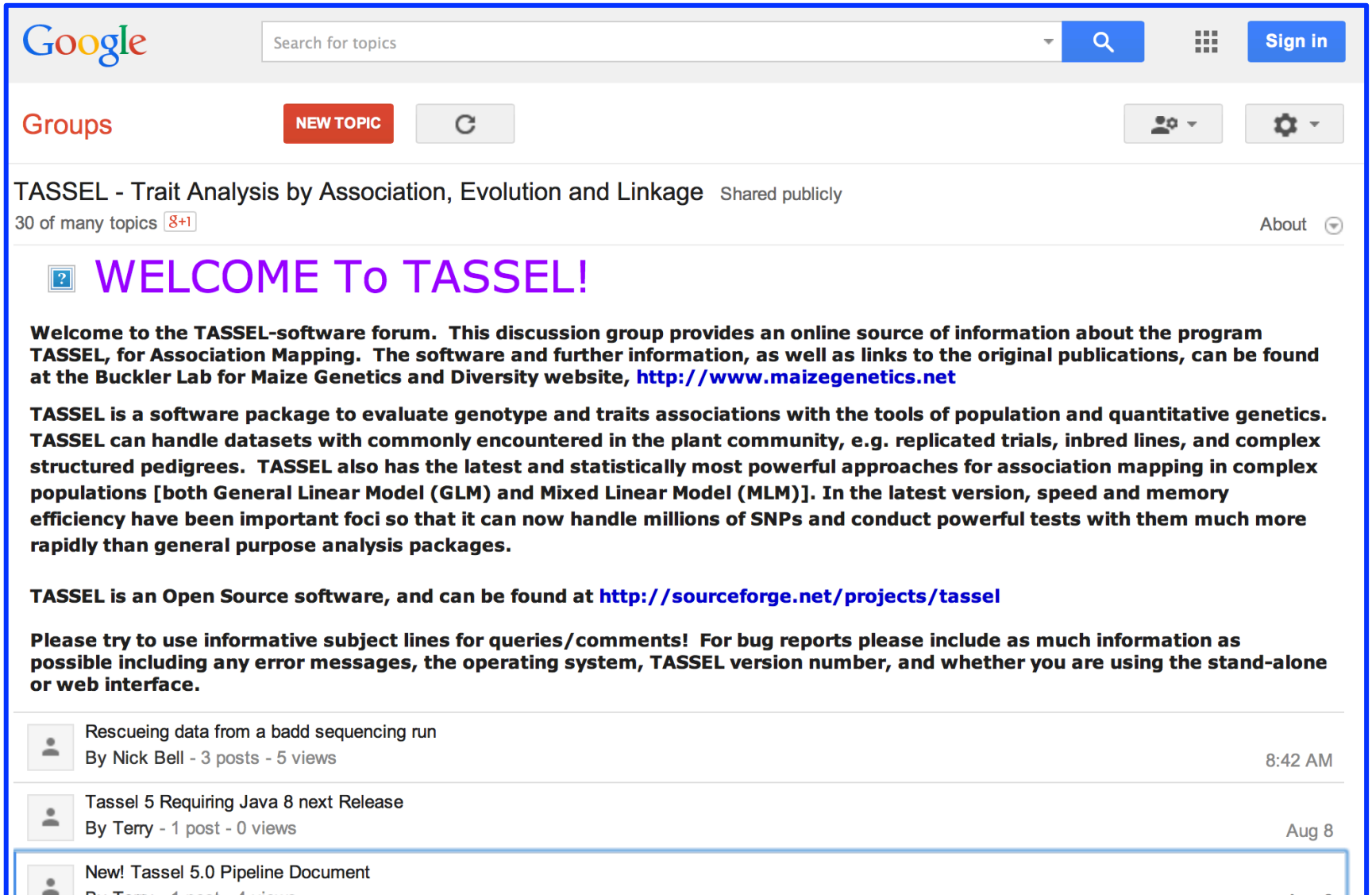
TASSEL 5.0 Pipeline Command Line Interface:
Guide to using Tassel Pipeline

Terry Casstevens (*tmc46@cornell.edu*)
Institute for Genomic Diversity, Cornell University, Ithaca, NY 14853-2703
April 9, 2015

Prerequisites

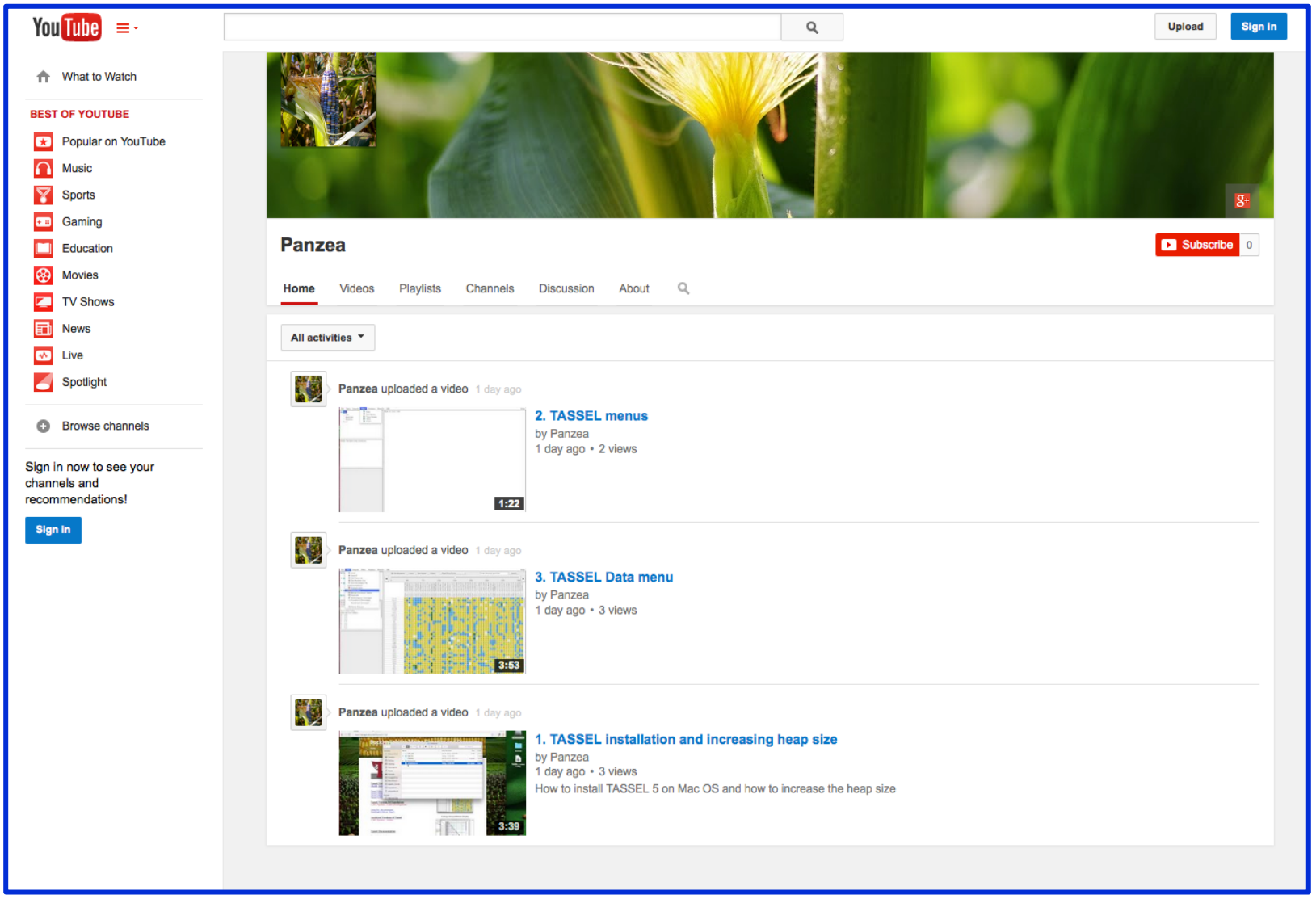
Resources for Tassel Users

- Available via www.maizegenetics.net/tassel

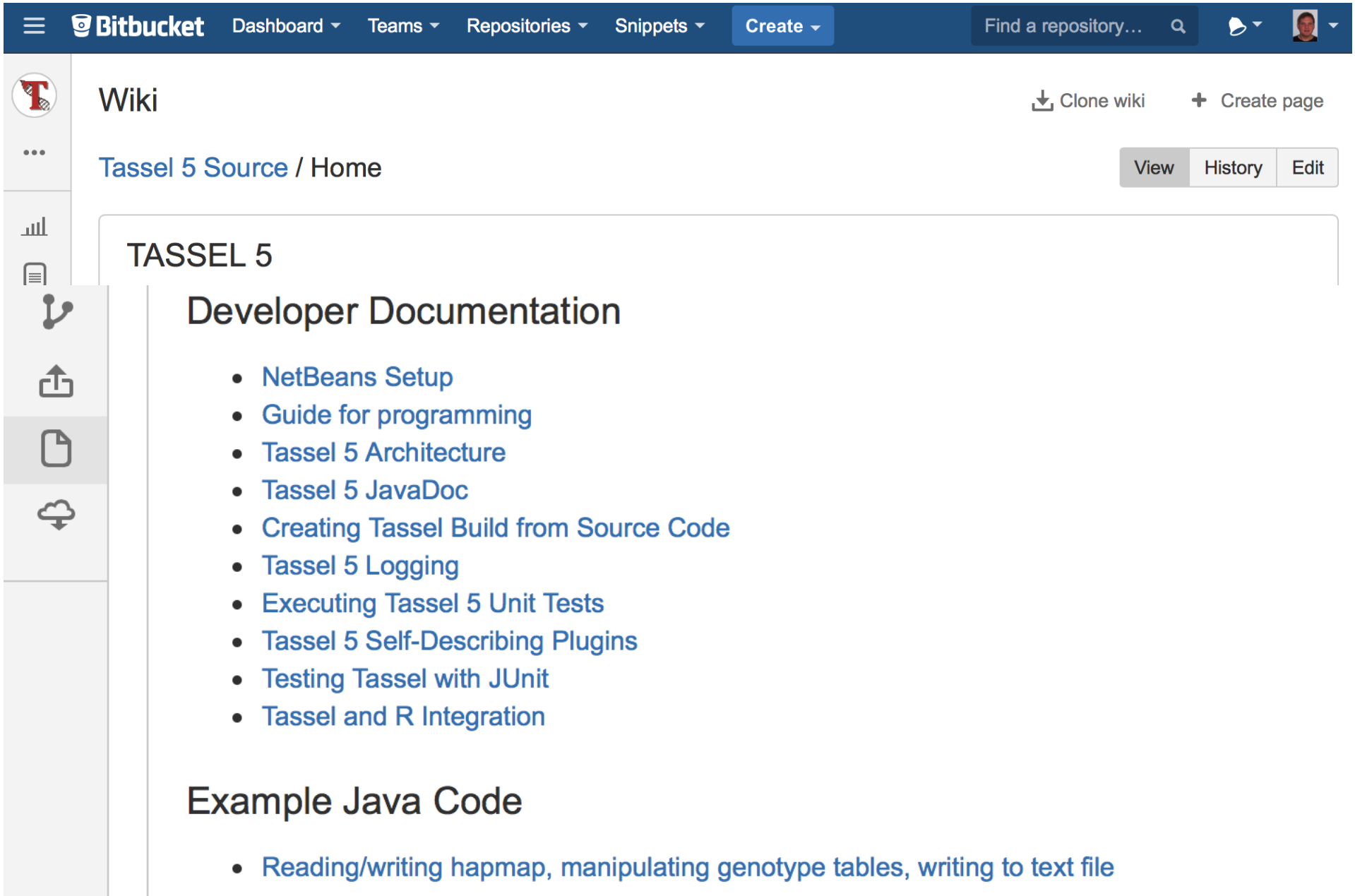


Resources for Tassel Users

- Available via www.maizegenetics.net/tassel



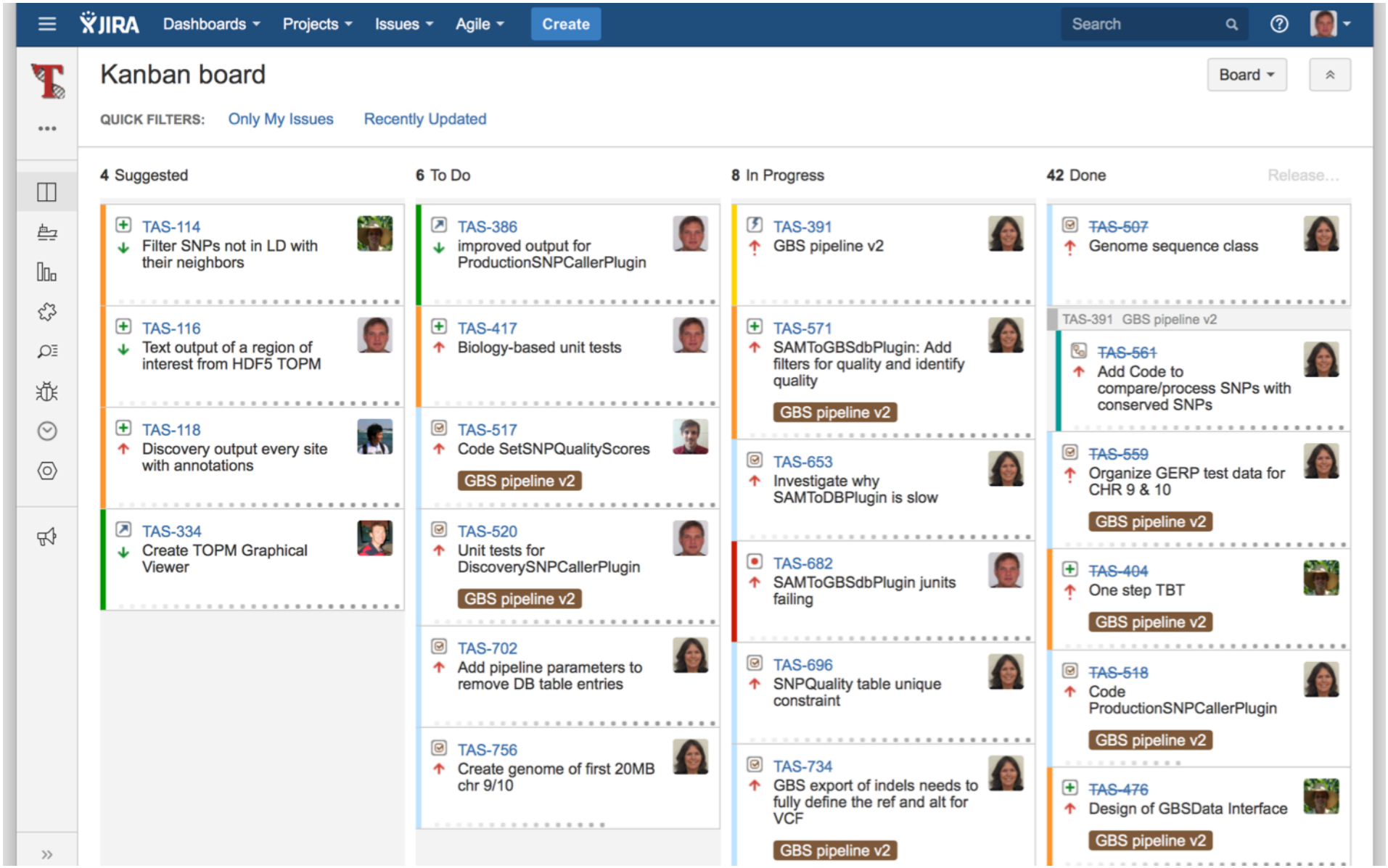
Resources for Tassel Developers



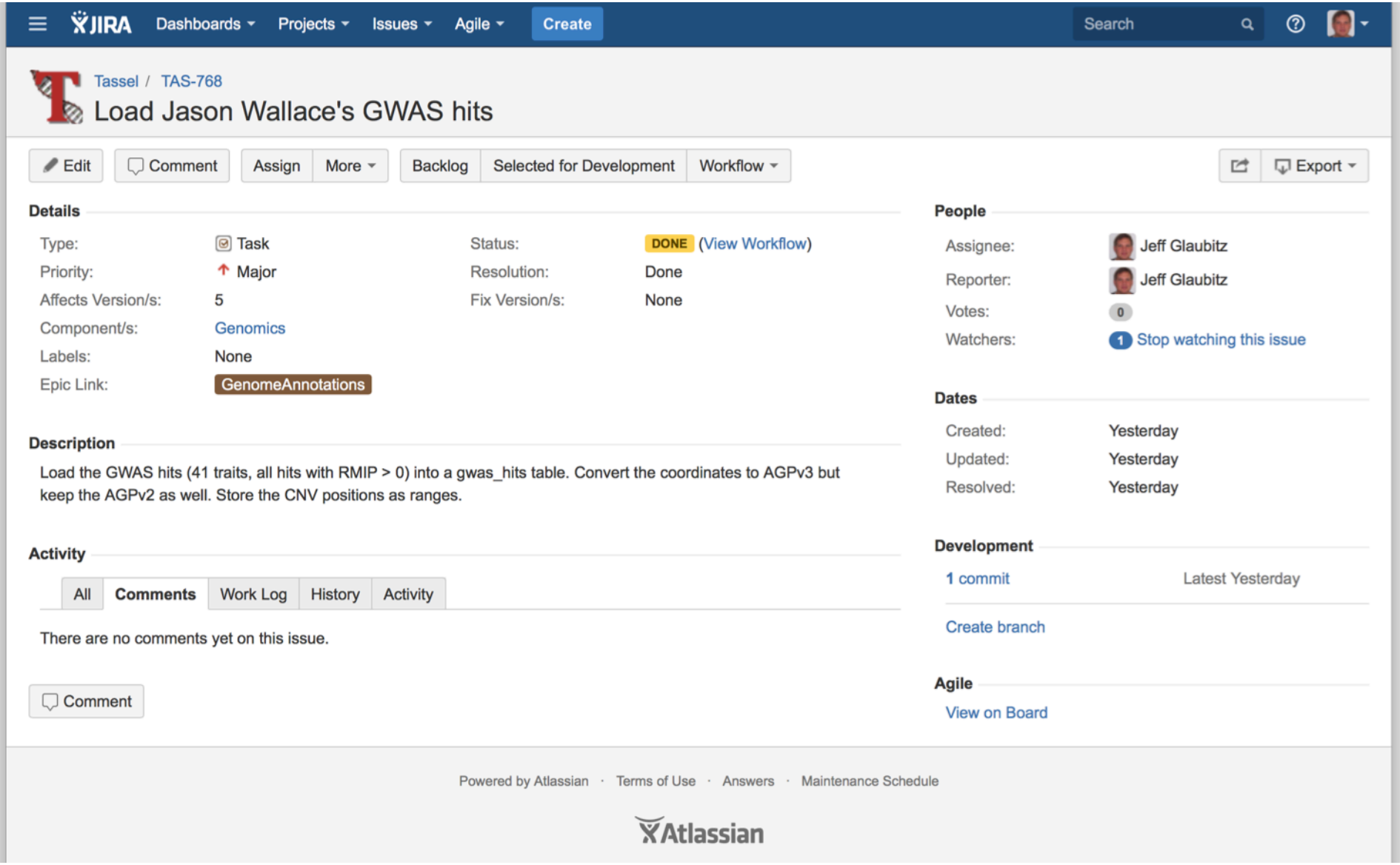
Resources for Tassel Developers

- Tassel Hackathon: May 4-8, 2015
  - Keep in mind for next year!
- Useful tools
  - IDEs: **NetBeans**, **IntelliJ**, **Eclipse**
  - Version control: **git** (via **Tower** on Mac)
  - Source code repository: **BitBucket**
  - Task management: **JIRA**
    - Integrates with BitBucket

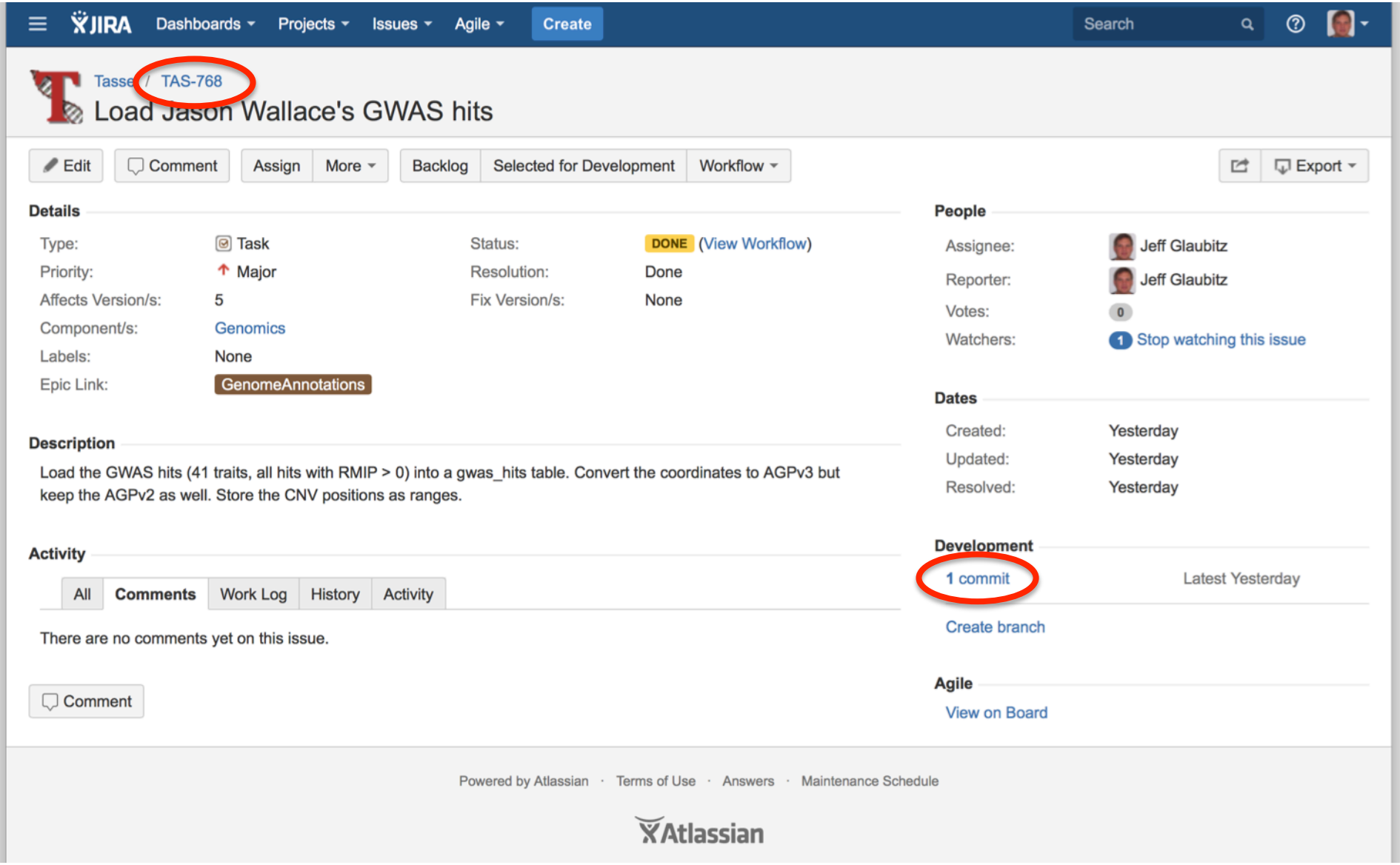
JIRA “Kanban board”



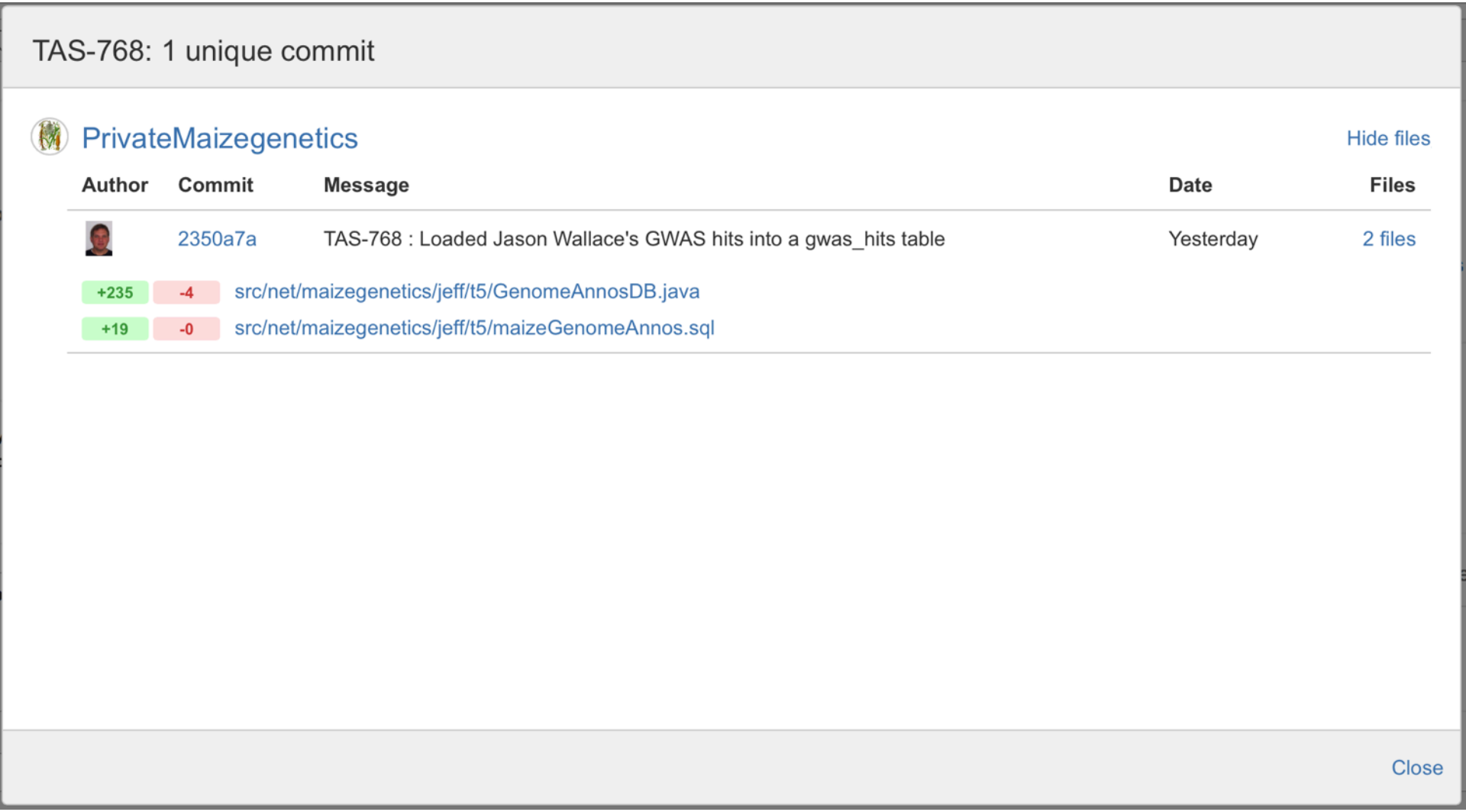
JIRA Issue


JIRA
Dashboards
Projects
Issues
Agile
Create
Search
Tassel / TAS-768
Load Jason Wallace's GWAS hits
Edit
Comment
Assign
More
Backlog
Selected for Development
Workflow
Export
Details
Type: Task
Priority: Major
Affects Version/s: 5
Component/s: Genomics
Labels: None
Epic Link: GenomeAnnotations
Status: DONE (View Workflow)
Resolution: Done
Fix Version/s: None
Description
Load the GWAS hits (41 traits, all hits with RMIP > 0) into a gwas_hits table. Convert the coordinates to AGPv3 but keep the AGPv2 as well. Store the CNV positions as ranges.
Activity
All
Comments
Work Log
History
Activity
There are no comments yet on this issue.
Comment
People
Assignee: Jeff Glaubitz
Reporter: Jeff Glaubitz
Votes: 0
Watchers: 1 Stop watching this issue
Dates
Created: Yesterday
Updated: Yesterday
Resolved: Yesterday
Development
1 commit
Latest Yesterday
Create branch
Agile
View on Board
Powered by Atlassian · Terms of Use · Answers · Maintenance Schedule
Atlassian

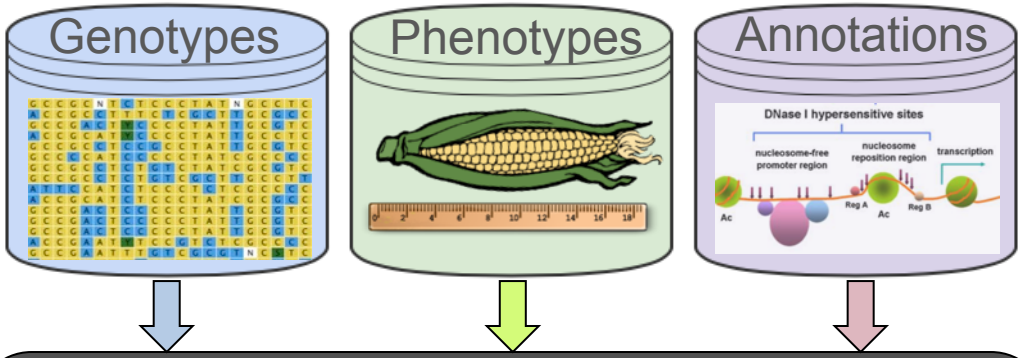
Bitbucket/JIRA Integration



Bitbucket/JIRA Integration



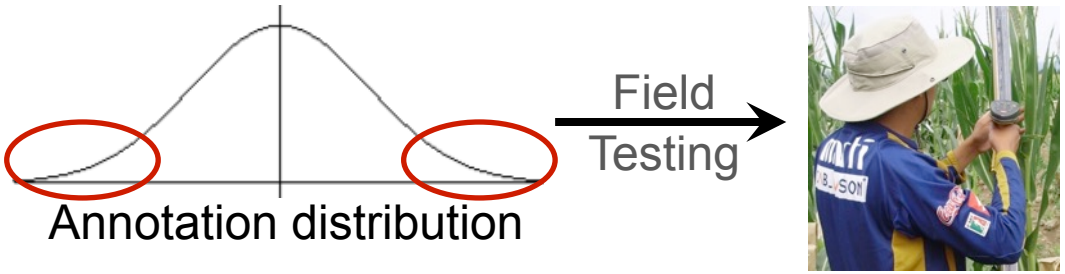
Data:



Analysis:

GWAS
Machine Learning
Variance (h^2) Explained
Enhanced Genomic Prediction

Verification:

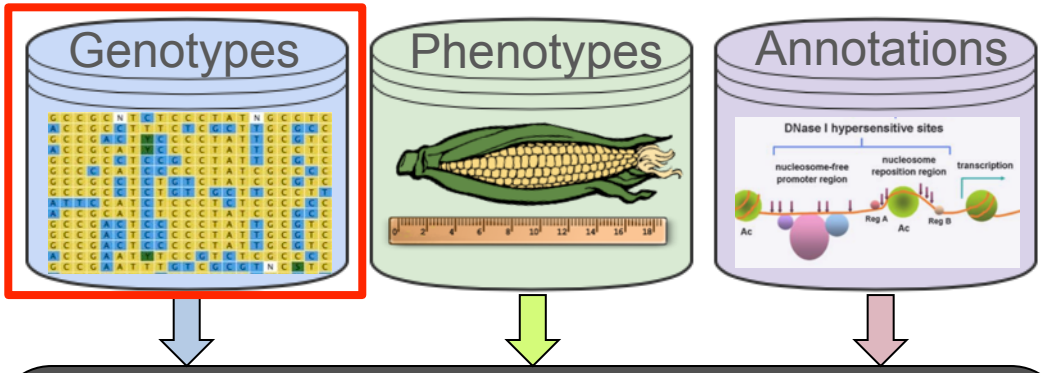


Products:

Accelerated Breeding

Personalized Medicine for Corn?

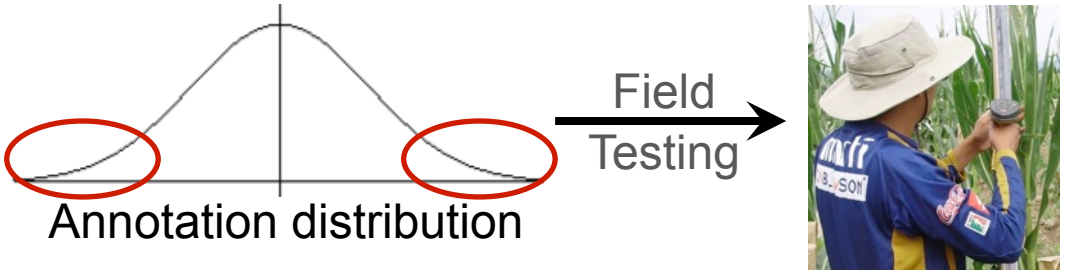
Data:



Analysis:

GWAS
Machine Learning
Variance (h^2) Explained
Enhanced Genomic Prediction

Verification:

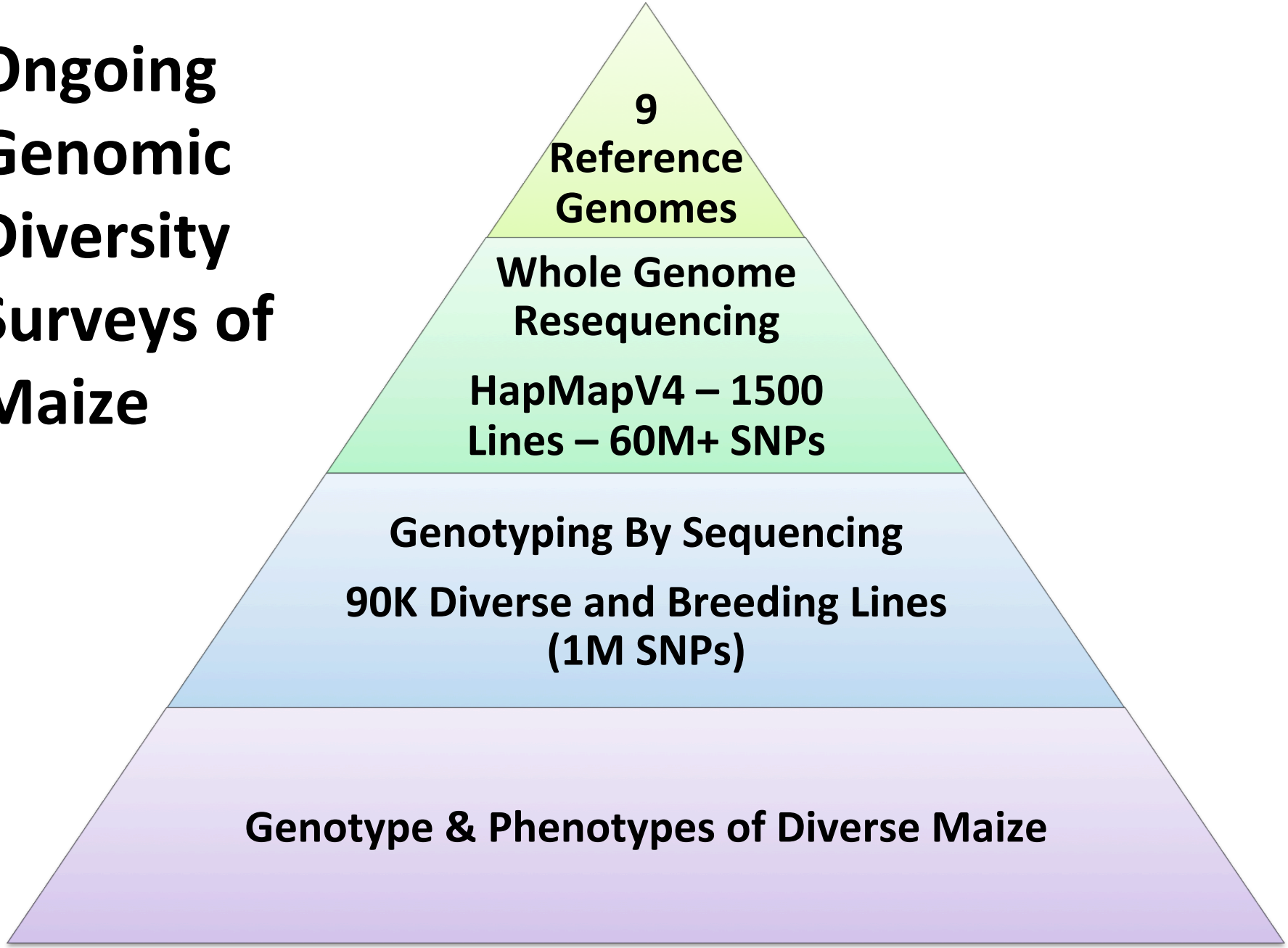


Products:

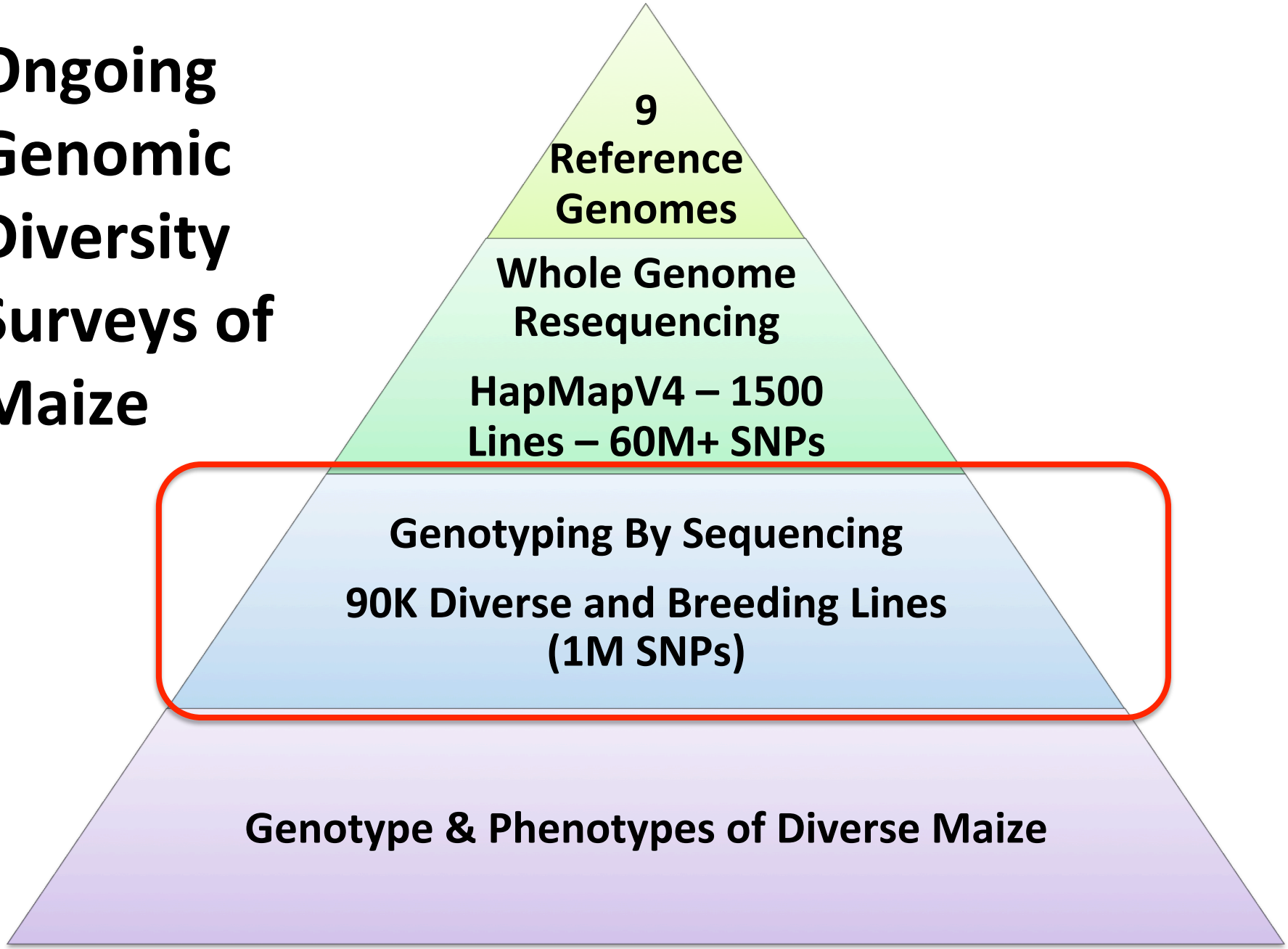
Accelerated Breeding

Personalized Medicine for Corn?

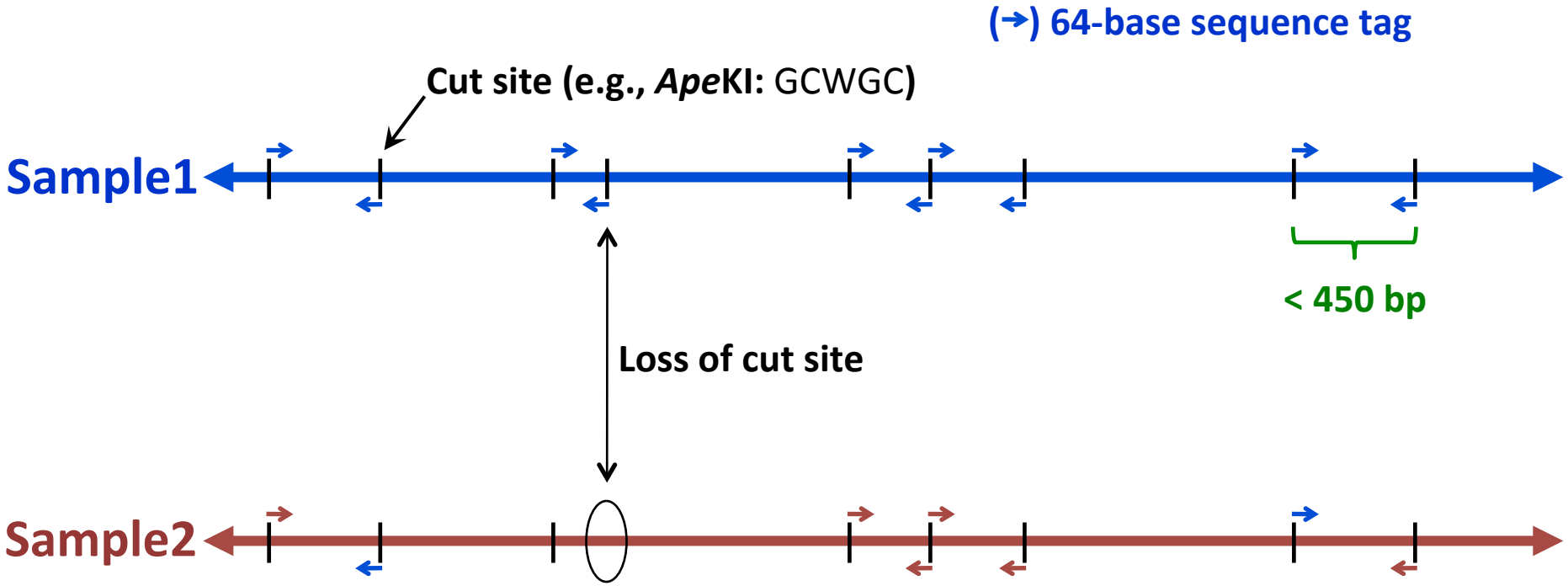
Ongoing Genomic Diversity Surveys of Maize



Ongoing Genomic Diversity Surveys of Maize



What is genotyping-by-sequencing (GBS)?

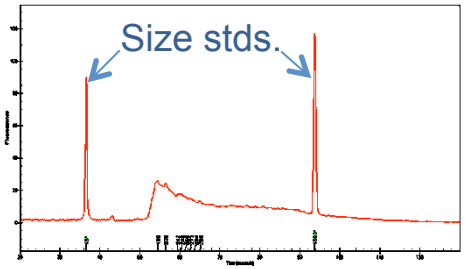


- Reduced representation approach inspired by Altshuler *et al.* (2000)
- Focuses NextGen sequencing power to ends of restriction fragments
- Scores both biallelic markers and presence/absence markers

The GBS protocol is simple and robust

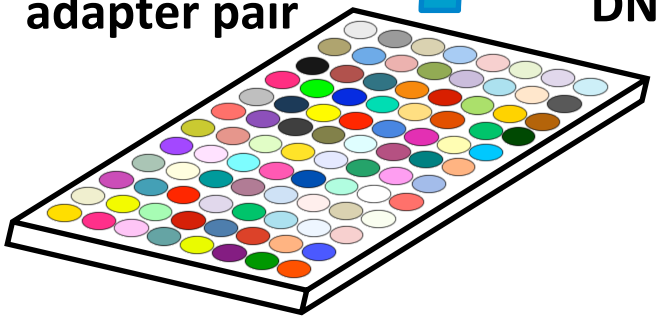
Elshire *et al.* 2011. PLoS ONE

Rob Elshire

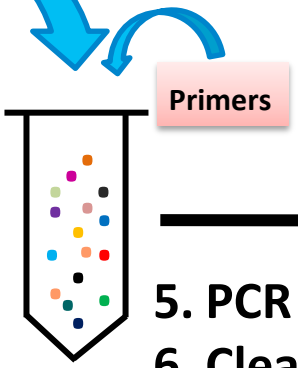


7. Evaluate fragment sizes

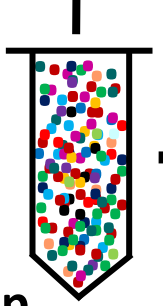
1. Plate DNA & adapter pair



4. Pool DNAs



5. PCR
6. Cleanup



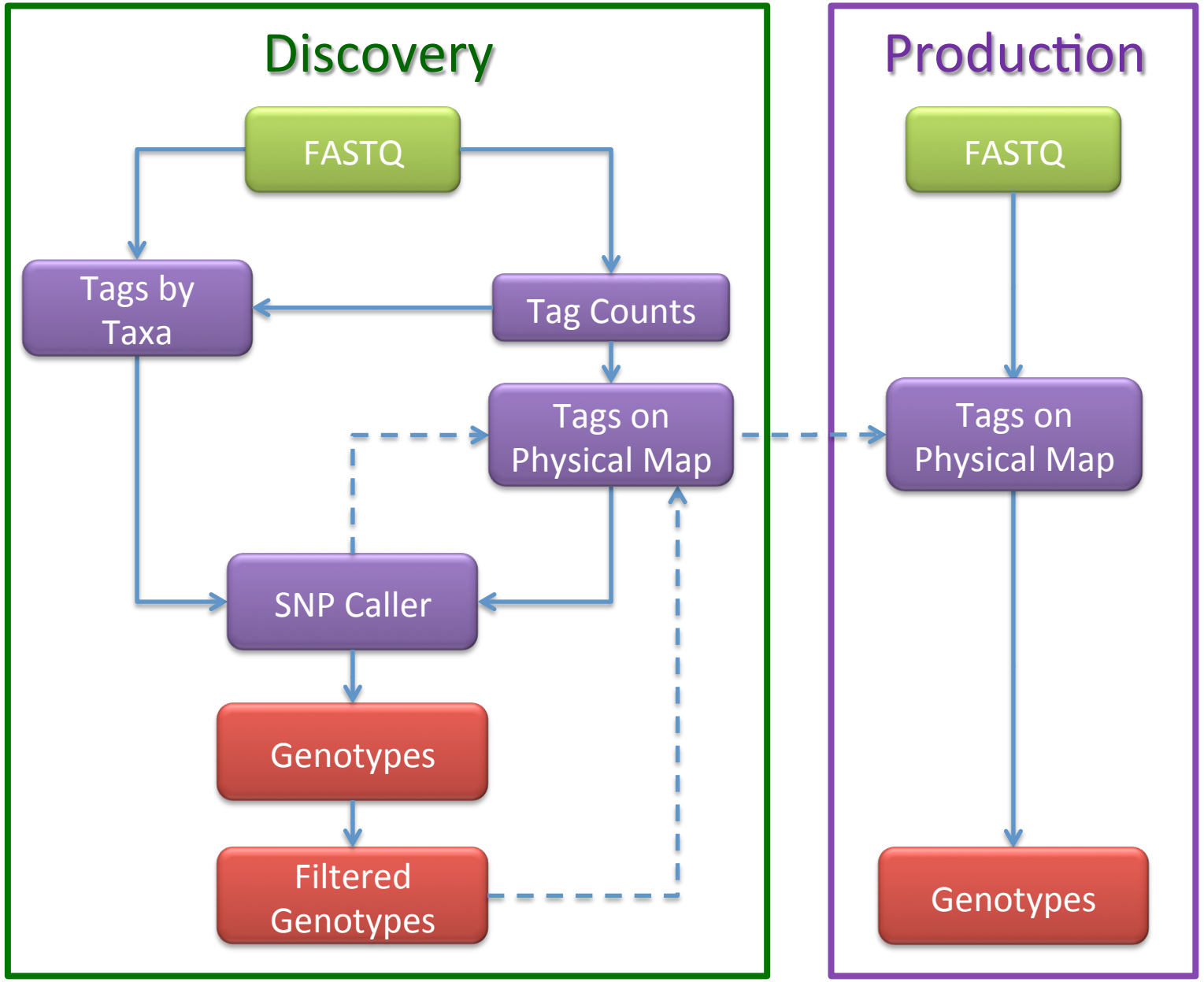
2. Digest DNA with RE (e.g. *Ape*KI)
3. Ligate adapters

8. Run on single lane of Illumina flowcell

Sequence Reads in FASTQ format

Reference-based GBS bioinformatics pipeline

Glaubitz *et al.* 2014 PLoS ONE

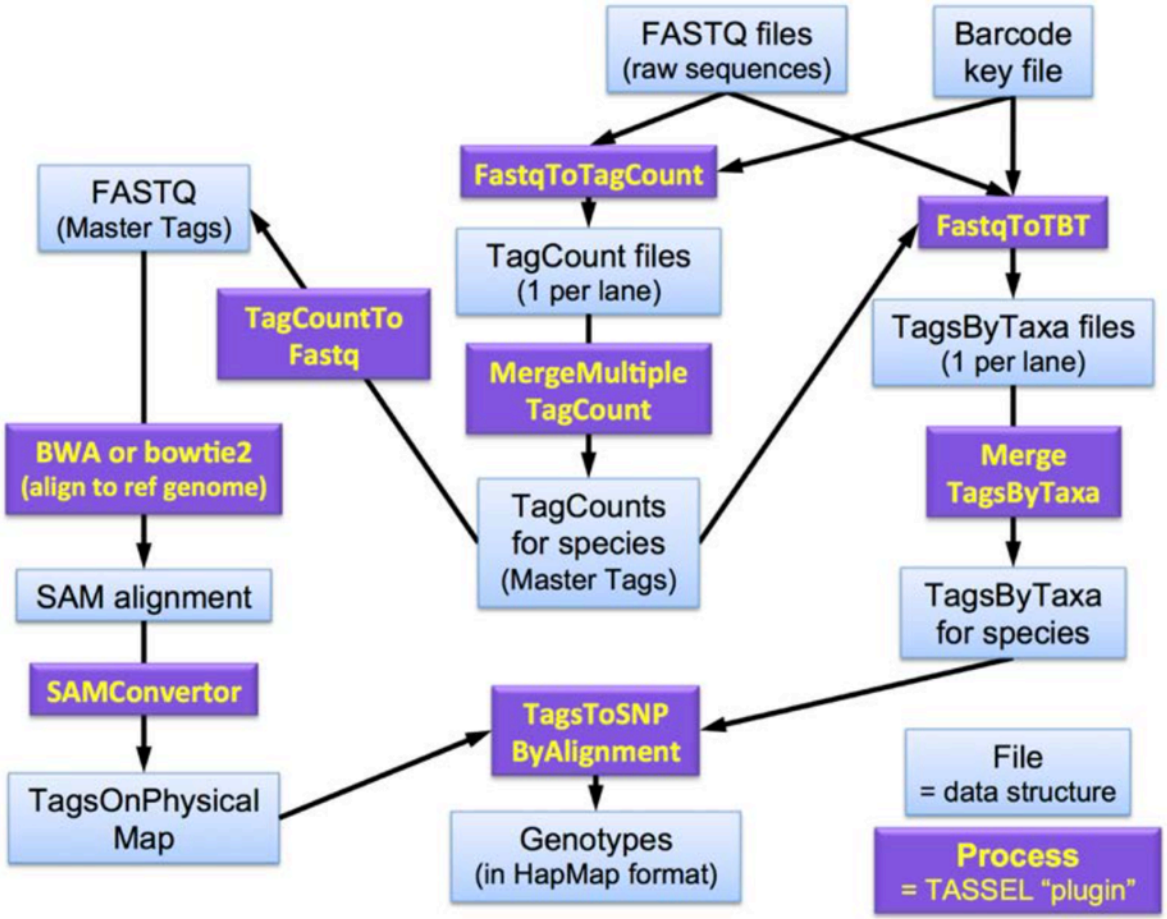


Ed Buckler

Jeff Glaubitz

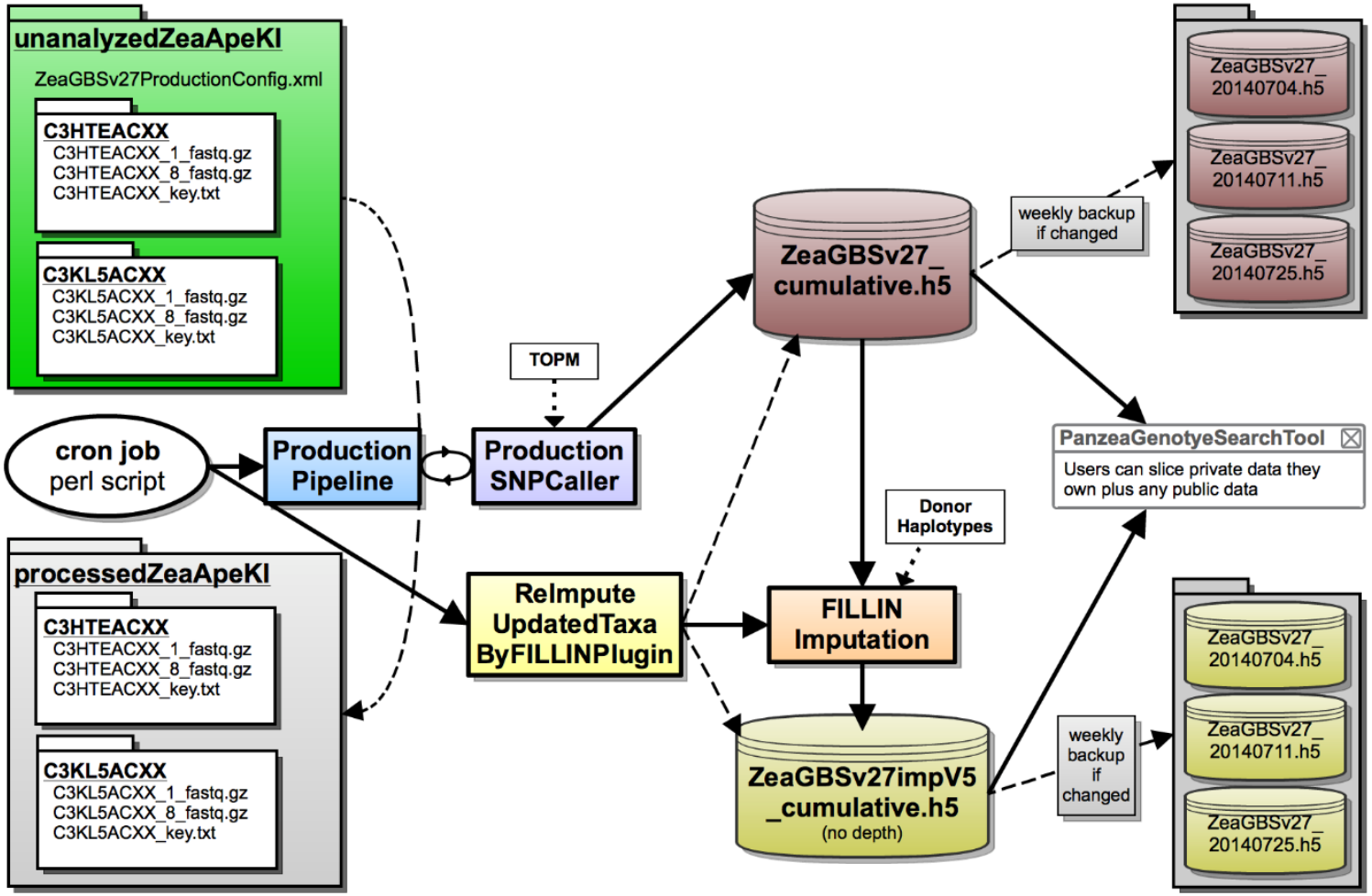
Terry Casstevens

TASSEL-GBSv1 Discovery Pipeline details



Automated GBSv1 Production Pipeline for Maize

- Improved data turnabout time with weekly build
- Improved scalability (currently at >70K taxa by 1M SNPs)



GBS v1 Problems

1. **Sequential processing of flowcells for discovery**
   - 30-40 minutes each
   - 175 hours to process 300 flowcells
2. **Custom data formats for all files**
3. **Fixed length sequences**
4. **Quality scores ignored**
5. **Data structures all favor space, not speed**

GBS v2 Design


Ed Buckler

1. Optimized for speed
   - Requires 100 to 500Gb machine for maize build
   - New compression algorithms to hold taxa distributions
2. Relational Database at the heart
   - Allows for much more complex queries
3. Clean objects defining the data structures to work with TASSEL 5
4. Quality scores and length considered
5. User focused scoring approaches

GBSv2 Speed Gains

1. **Parallel fastq file processing**
   - **30-40X faster with 64-core machine**
   - **Hashmap used for tag lookup (18X)**
   - **Trie used for barcode processing (Janu Verma)**
2. **Parallel SNP calling**
   - **SNP calling should be able to scale with cores**

GBSv2 Flow


Ed Buckler

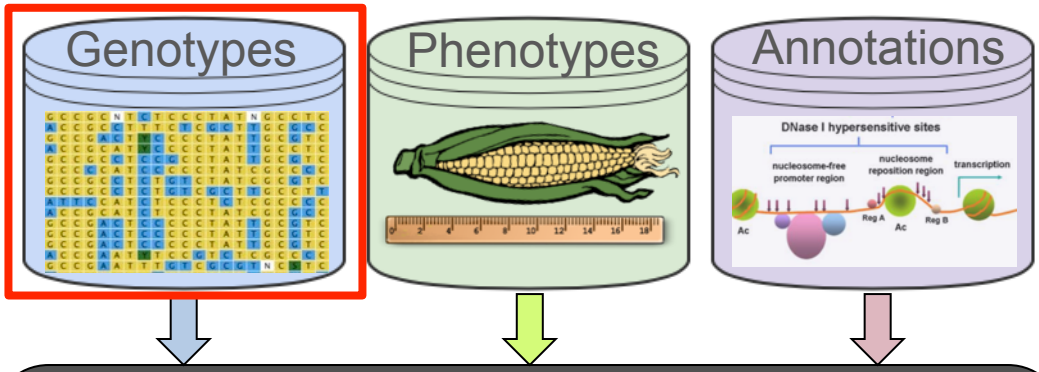

Lynn Johnson


Zak Miller

Which SNPs to keep?

1. **Pipeline can characterizes SNP distribution for custom lists of taxa**
   - **Calculates many of the statistics only when depth**
2. **Tools for comparing SNP overlap**
   - **HapMap or GERP pipelines can be used to provide false positive and negative rates**
3. **Invariant sites can be output**
   - **Key for popgen and deleterious mutation studies**

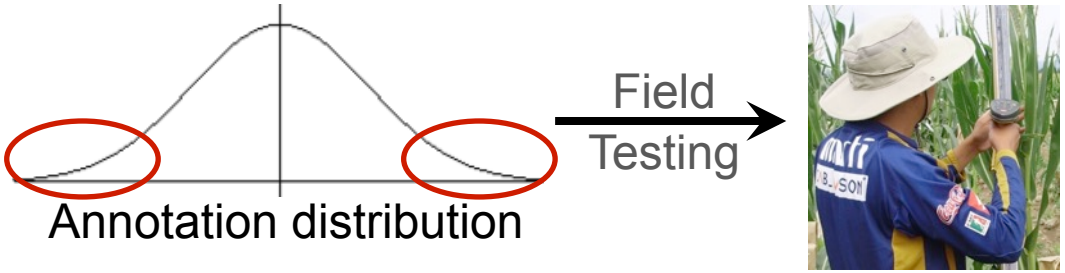
Data:



Analysis:

GWAS
Machine Learning
Variance (h^2) Explained
Enhanced Genomic Prediction

Verification:

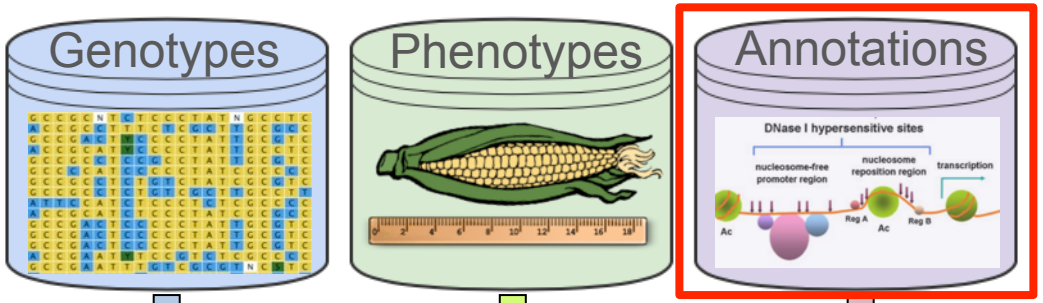


Products:

Accelerated Breeding

Personalized Medicine for Corn?

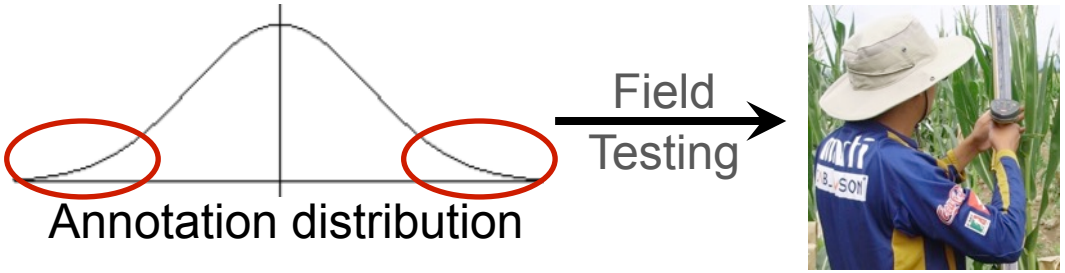
Data:



Analysis:

GWAS
Machine Learning
Variance (h^2) Explained
Enhanced Genomic Prediction

Verification:



Products:

Accelerated Breeding

Personalized Medicine for Corn?

Genome Annotations DB

- Implemented in PostgreSQL 

  - Populated with Java code
- Schema currently has 4 modules:
  - Genome (genes, transcripts, exons, CDS, AGPv2 to v3)
  - GWAS hits and QTL peaks
  - Annotation Names (associated metadata)
  - Site Annotations (a single, large table)
    - Currently holds annotations for 60M variable sites
- Shared with project members via private folder on iPlant
  - Documentation on how to install & query
  - Formats for contributing new annotations
  - Monthly (private) releases on iPlant

Scaling to the whole genome?

- Challenge for the next year
- 2.3 billion sites vs. current 60 million (= 38x)
- Strategies to explore
  - Focus on important part of the genome
  - Store ranges instead of sites wherever possible
  - JSON fields in site annotations table
  - Columnar storage (e.g., CitusDB)
  - Blobs (bit storage)
  - FastBit (good for sparse data)
  - HDF5 (read only)
  - “Sharding” across multiple servers (CitusDB, MongoDB)