

# **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
- Accelerated 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

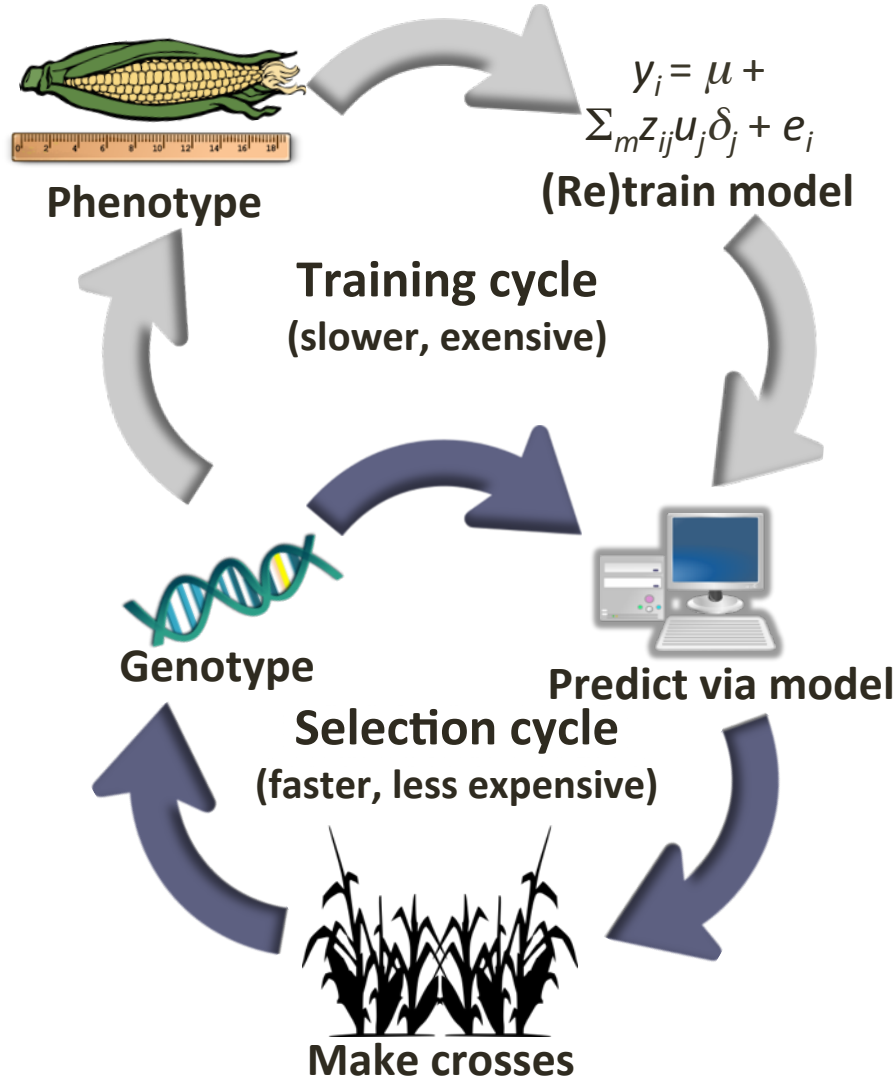
# Biology of rare alleles in maize and its wild relatives

## 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



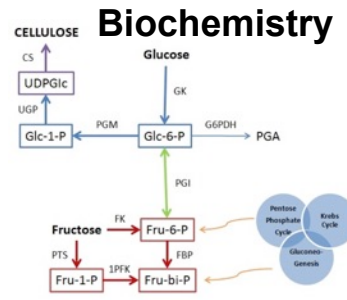
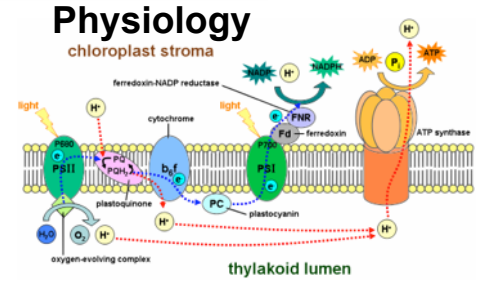
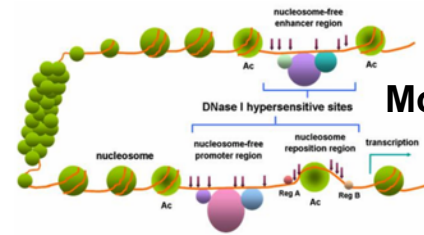
# Biology Assisted Breeding



Phenotype

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

(Re)train model



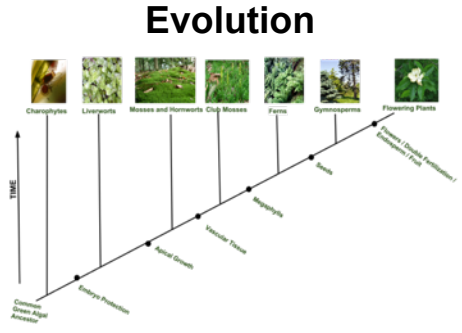
Training cycle  
 (slower, expensive)

Genotype  
 Selection cycle  
 (faster, less expensive)

Predict via model



Make crosses



Environment

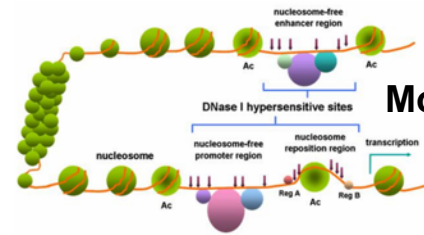
# Biology Assisted Breeding



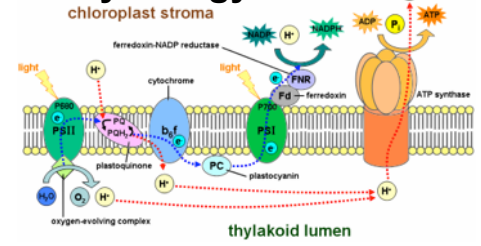
Phenotype

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

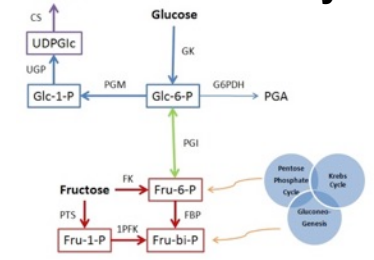
(Re)train model



**Physiology**



**Biochemistry**



Training cycle  
(slower, expensive)

Genotype  
Selection cycle  
(faster, less expensive)

Predict via model



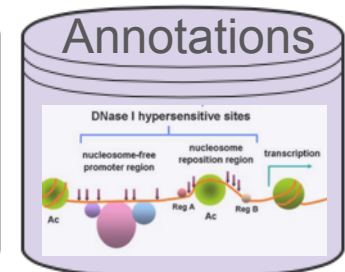
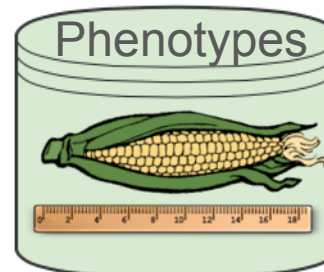
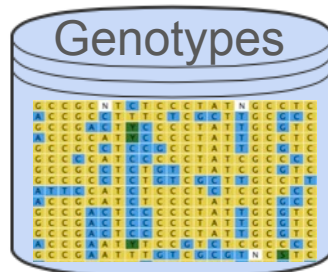
CRISPR

**Evolution**

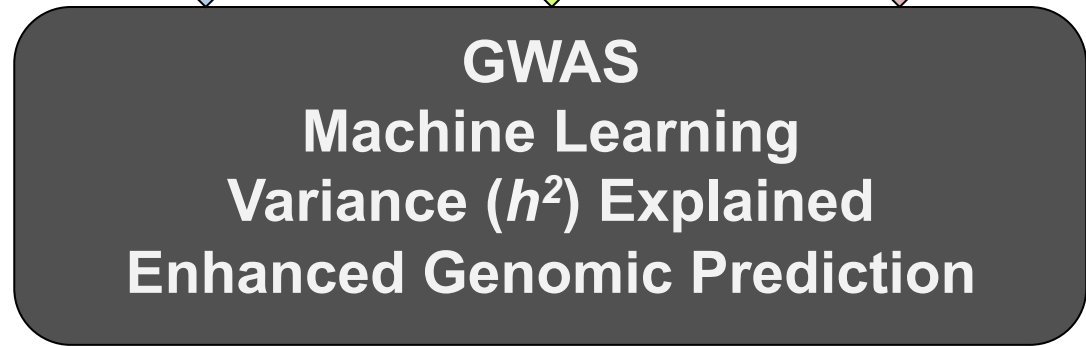


**Environment**

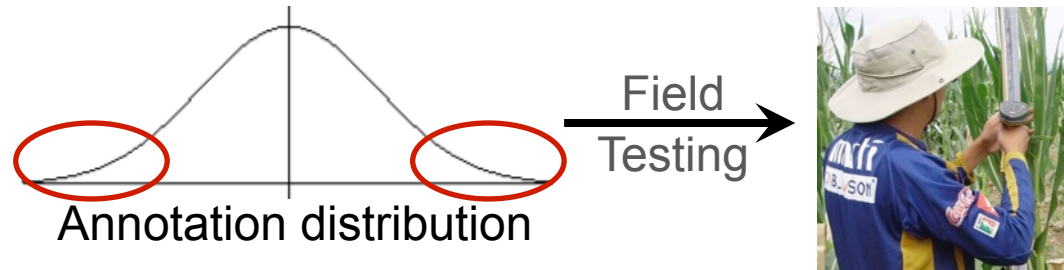
**Data:**



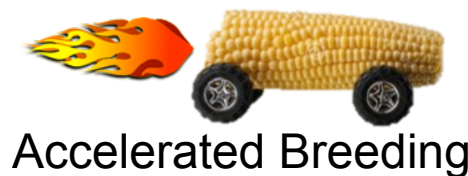
**Analysis:**



**Verification:**

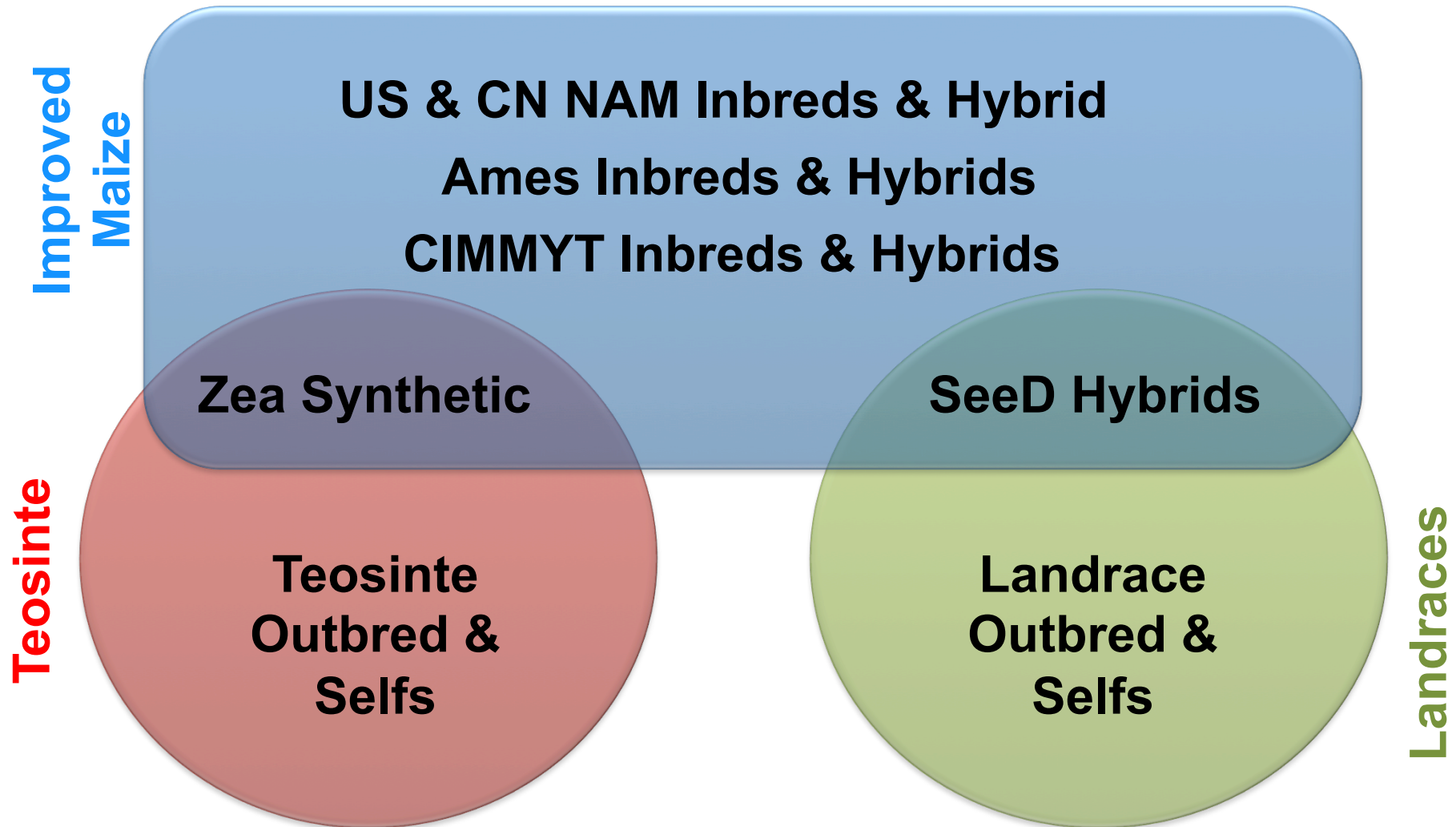


**Products:**

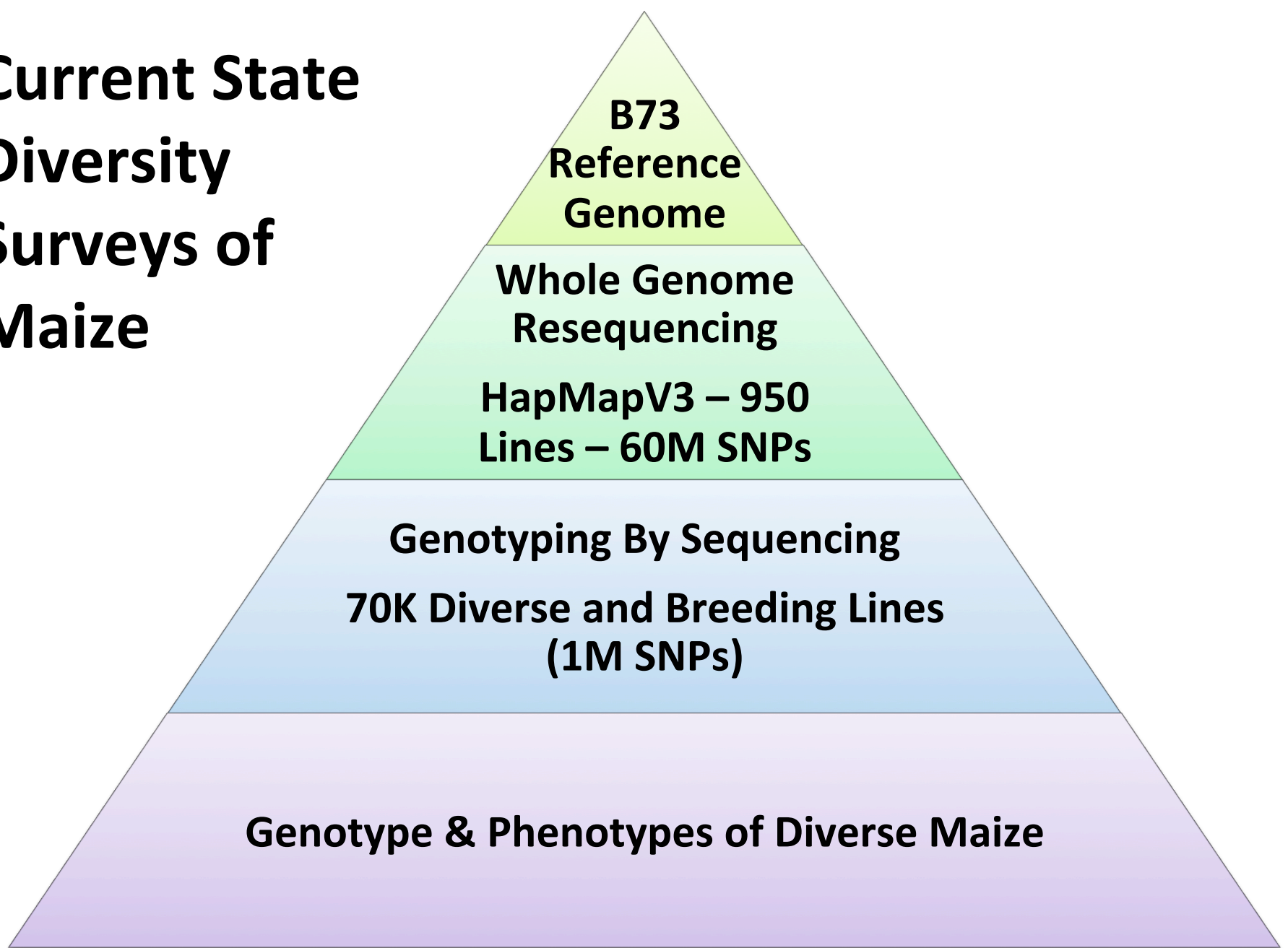


Personalized  
Medicine for  
Corn?

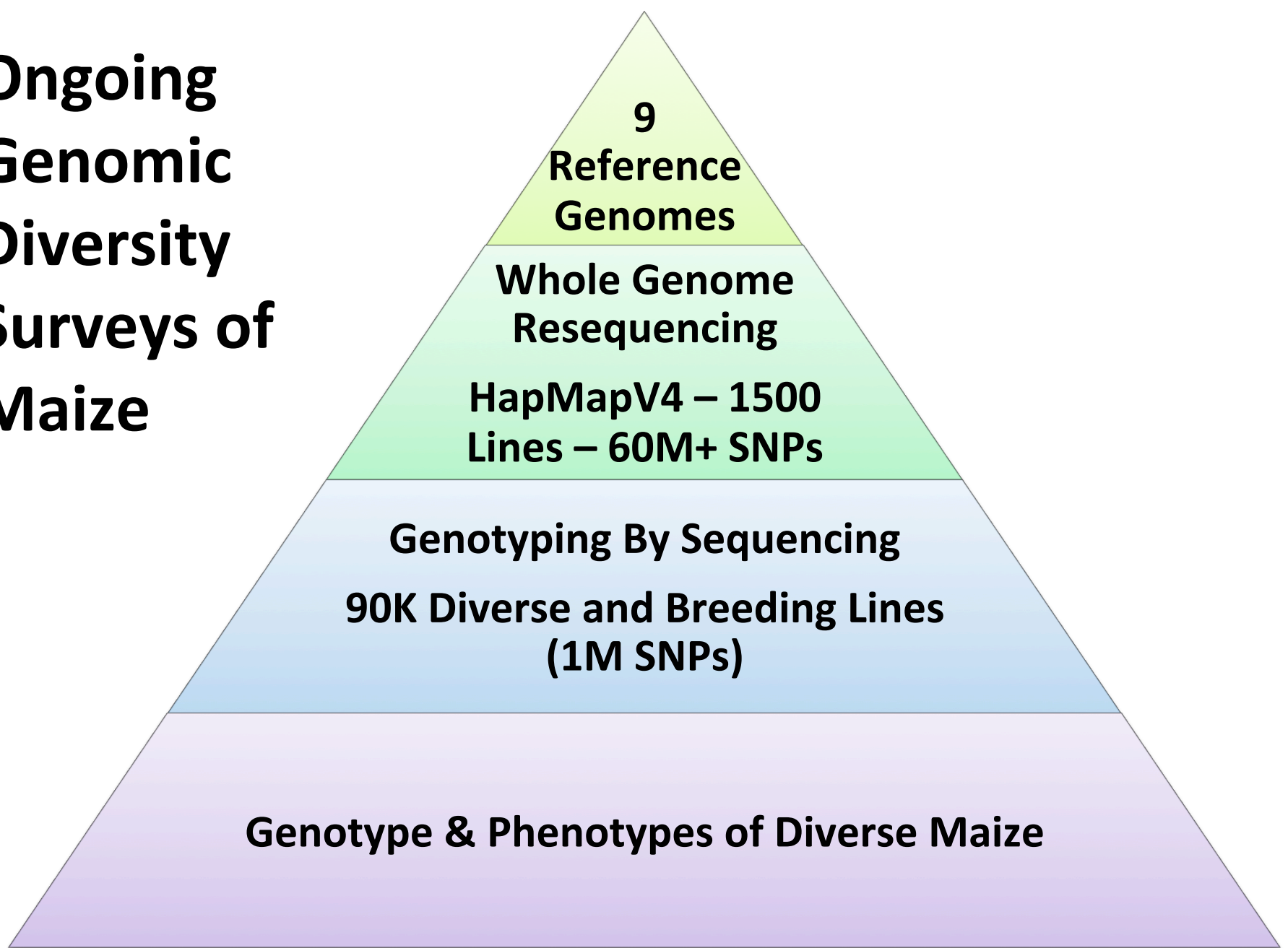
# The Genomic & Field Data Sample a Range of Historical Inbreeding with a Variety of Dominance Relationships



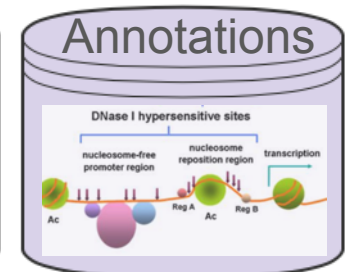
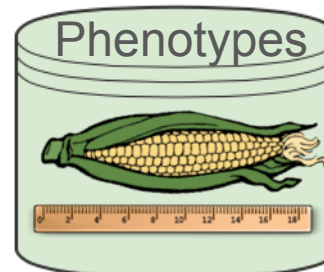
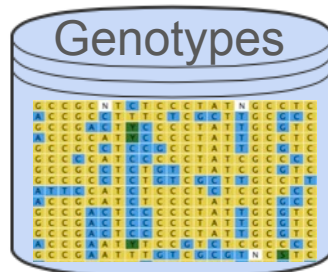
# Current State Diversity Surveys of Maize



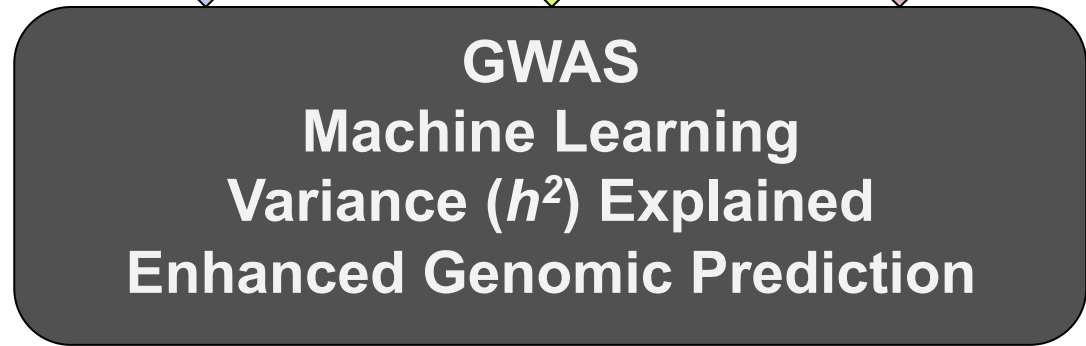
# Ongoing Genomic Diversity Surveys of Maize



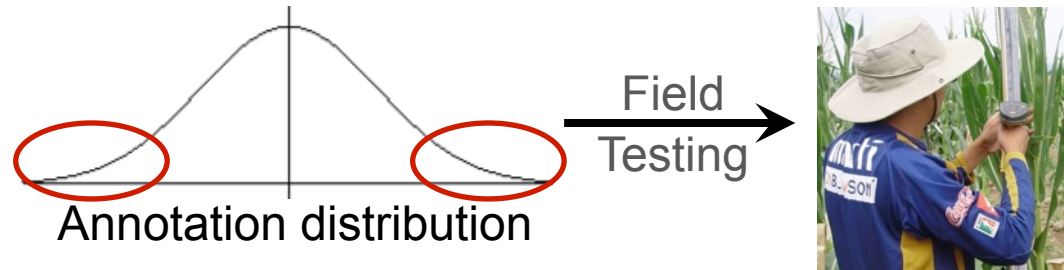
**Data:**



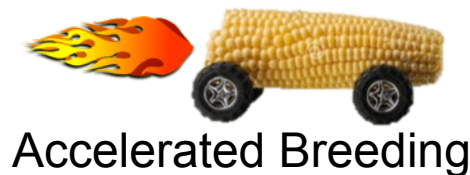
**Analysis:**



**Verification:**



**Products:**



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**



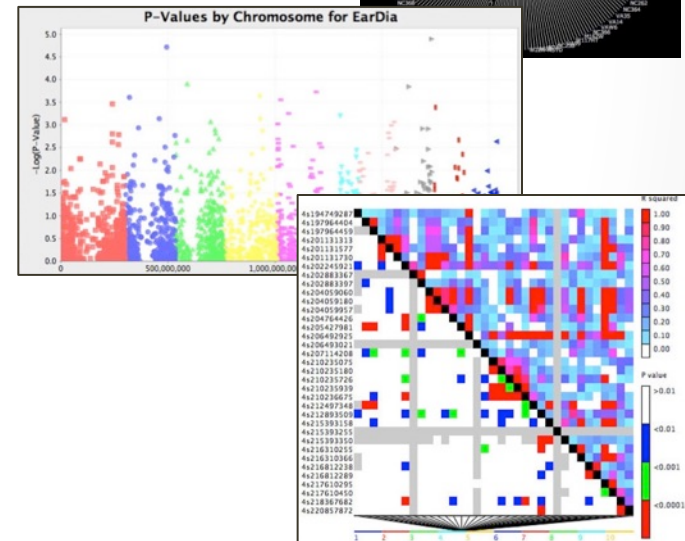
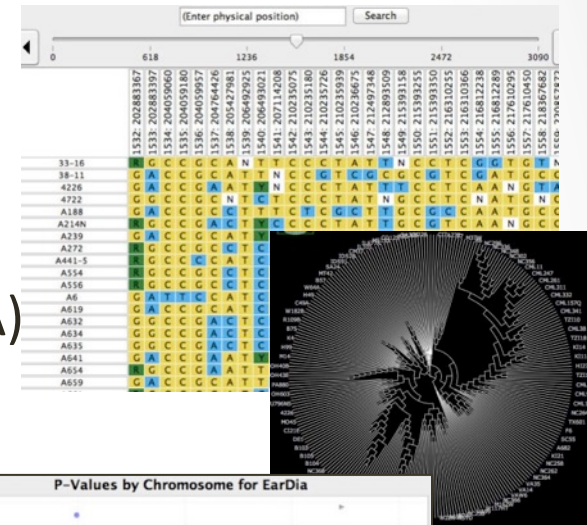
**Fei Lu**



**Jeff Glaubitz**

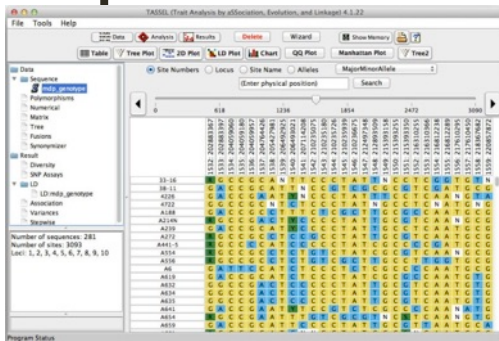
# What TASSEL Can & Can't Do

- 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);
```

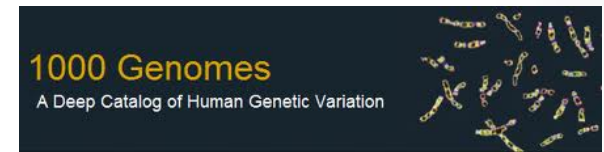
```
...
```

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



# General Philosophy - GUI

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

The screenshot shows the TASSEL (Trait Analysis by aSSociation, Evolution, and Linkage) 5.0.5 interface. The 'Data' menu is open, listing options such as Load, Export, Transform, Synonymizer, Intersect Join, Union Join, Merge Alignments, Separate, Homozygous Genotype, and Delete Dataset. The main window displays a table of genetic data with columns for site numbers and alleles. A search bar is visible above the table.

	165: 6755	166: 6795	167: 6835	168: 6918	169: 7000	170: 7040	171: 7155	172: 7190	173: 7229	174: 7267	175: 7345	176: 7465	177: 7584	178: 7663	179: 7699	180: 7735	181: 7808	182: 7888	183: 8240	184: 8311	185: 0	186: 166
07ril-100	A	C	C	C	A	C	G	A	T	A	T	T	A	T	A	T	T	G	T	G	A	G
07ril-106	A	C	C	C	A	C	G	A	T	A	T	T	A	T	A	T	T	G	T	G	A	G
07ril-121	C	T	T	G	G	G	A	G	C	G	A	C	G	C	T	C	A	C	C	A	A	G
07ril-126	C	T	T	G	G	G	A	G	C	G	A	C	G	C	T	C	A	C	C	A	A	G
07ril-127	C	T	T	G	G	G	A	G	C	G	A	C	G	C	T	C	A	C	C	A	A	G
07ril-129	A	C	C	C	A	C	G	A	T	A	T	T	A	T	A	T	T	G	T	G	G	C
07ril-130	C	T	T	G	G	G	A	G	C	G	A	C	G	C	T	C	A	C	T	G	G	C
07ril-131	C	T	T	G	G	G	A	G	C	G	A	C	G	C	T	C	A	C	C	A	A	G
07ril-132	C	T	T	G	G	G	A	G	C	G	A	C	G	C	T	C	A	C	C	A	A	G

# 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`

Perl script to start TASSEL

`-Xms2g -Xmx8g`

Java arguments to set memory

`-fork1`

(*Start of first command set*)

`-vcf myfile.vcf`

Load file

`-includeTaxaInFile 0_core_taxa.txt`

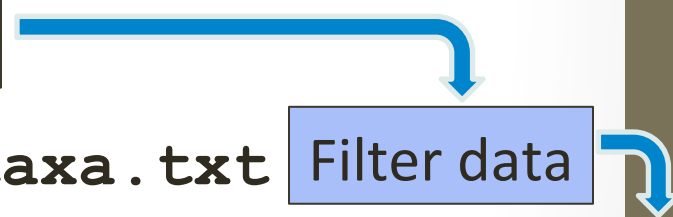
Filter data

`-export myfile_filtered.hmp.txt.gz`

Export result

`-runfork1`

*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 = text 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](http://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)*

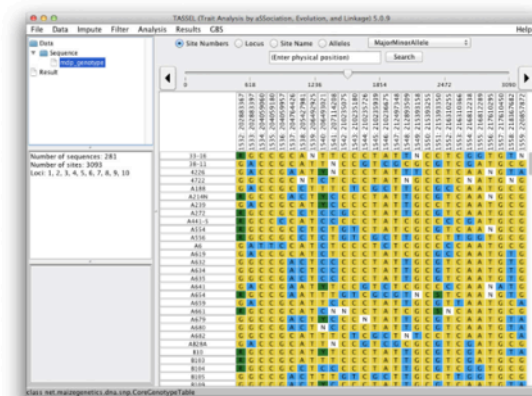
[Using Git - Recommended!](#)

[Download \(Click on "Tags"\)](#)

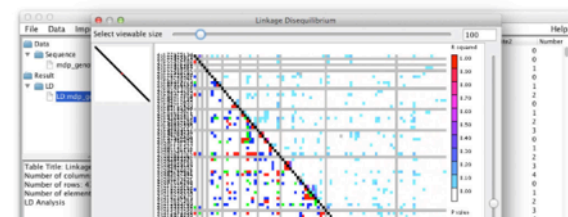
**Archived Versions of Tassel**  
*(GBS Pipeline - Stable)*

**Tassel Documentation**

*Alignment Viewer*



*Linkage Disequilibrium Display*



# Resources for Tassel Users

- Available via [www.maizegenetics.net/tassel](http://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](http://www.maizegenetics.net/tassel)



## Wiki

 Clone wiki  Create page

### Tassel 5 Source / Home

 View  History  Edit

## TASSEL 5

TASSEL project is a Java platform designed for the optimized analysis of crop genomic diversity.

### User Documentation

- [Reporting Tassel 5 Issues](#)
- [Tassel 5 User Manual](#)
- [Tassel 5 GBS v2 Pipeline](#)
- [Installing / Updating Tassel using Git \(\*recommended\*\)](#)
- [Executing Tassel](#)
- [Tassel 5.0 Pipeline \(\*Command Line Interface\*\)](#)
- [Tassel 3.0 and 4.0 Pipeline \(\*Command Line Interface\*\)](#)
- [Tassel GBS Pipeline](#)
- [Tassel UNEAK Pipeline](#)
- [Tassel Pan-genome Atlas \(PanA\) Pipeline](#)
- [Tassel Pipeline Tutorial](#)

### Developer Documentation

# Resources for Tassel Users

- Available via [www.maizegenetics.net/tassel](http://www.maizegenetics.net/tassel)

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

Terry Casstevens ([tmc46@cornell.edu](mailto:tmc46@cornell.edu))

Institute for Genomic Diversity, Cornell University, Ithaca, NY 14853-2703

*April 9, 2015*

[Prerequisites](#)

[Source Code](#)

[Install](#)

[Execute](#)

[Increasing Heap Size](#)

[Setting Logging to Debug or Standard \(With optional filename\)](#)

[Examples](#)

[Examples \(XML Configuration Files\)](#)

[Usage](#)

[Pipeline Controls](#)

[Data](#)

[Filter](#)

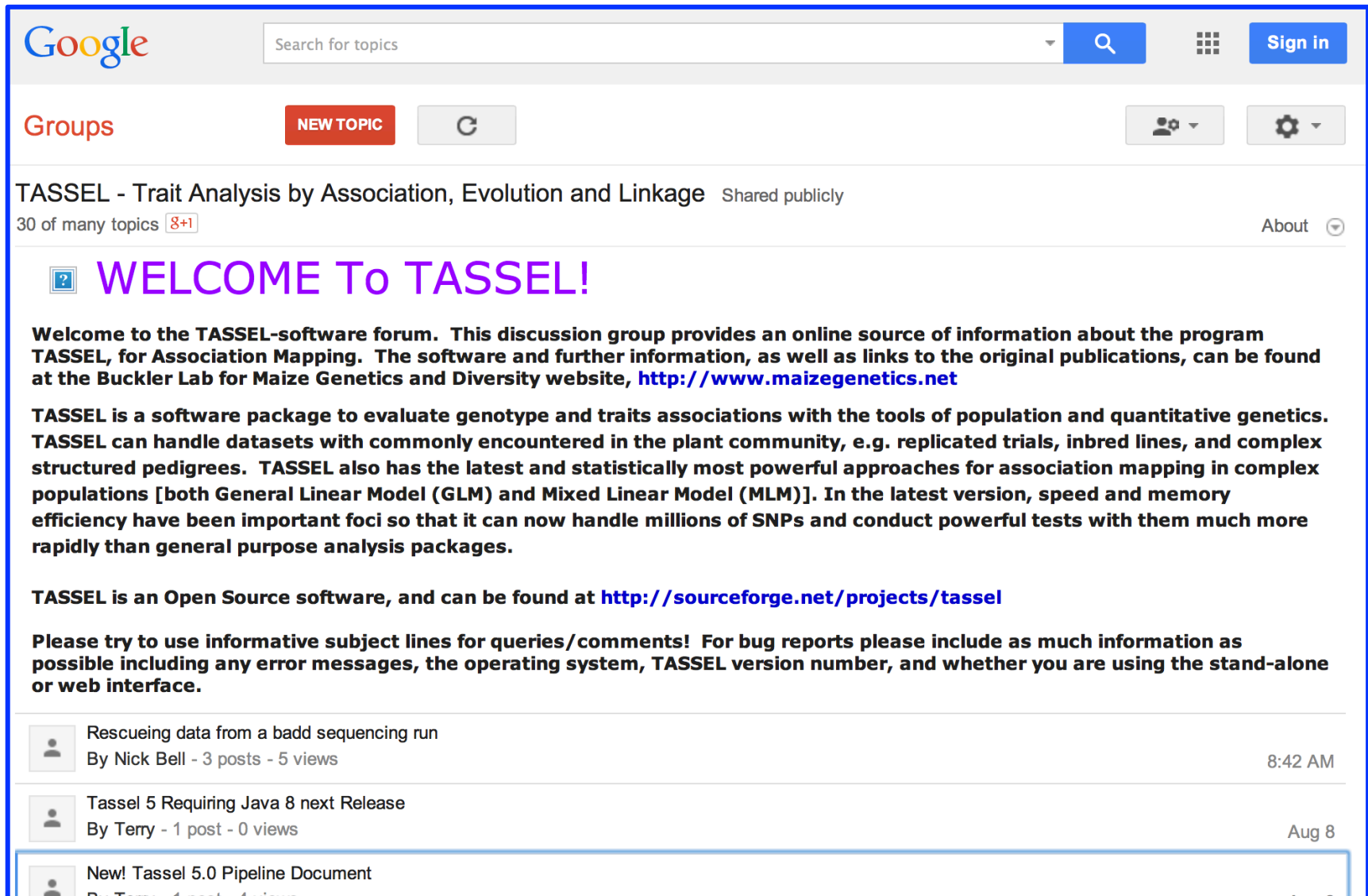
[Analysis](#)

[Results](#)

Prerequisites

# Resources for Tassel Users

- Available via [www.maizegenetics.net/tassel](http://www.maizegenetics.net/tassel)



The screenshot shows a Google Groups forum page. At the top, there is a Google search bar with the text "Search for topics" and a "Sign in" button. Below the search bar, the word "Groups" is displayed in red, followed by a "NEW TOPIC" button and a refresh icon. To the right, there are icons for user management and settings. The main heading of the group is "TASSEL - Trait Analysis by Association, Evolution and Linkage" with the status "Shared publicly". Below the heading, it says "30 of many topics" with a "+1" icon and an "About" link. The main content of the forum post is a purple heading "WELCOME To TASSEL!". Below this, there are three paragraphs of text: the first paragraph welcomes users and provides a link to the Buckler Lab website; the second paragraph describes the Tassel software's capabilities; the third paragraph states that Tassel is open source and provides a link to its SourceForge page. At the bottom, there are three forum posts listed, each with a user icon, the post title, the author's name, the number of posts and views, and the time or date.

Google Search for topics Sign in

Groups NEW TOPIC

TASSEL - Trait Analysis by Association, Evolution and Linkage Shared publicly

30 of many topics 8+1 About

## WELCOME To TASSEL!

Welcome to the TASSEL-software forum. This discussion group provides an online source of information about the program TASSEL, for Association Mapping. The software and further information, as well as links to the original publications, can be found at the Buckler Lab for Maize Genetics and Diversity website, <http://www.maizegenetics.net>

TASSEL is a software package to evaluate genotype and traits associations with the tools of population and quantitative genetics. TASSEL can handle datasets with commonly encountered in the plant community, e.g. replicated trials, inbred lines, and complex structured pedigrees. TASSEL also has the latest and statistically most powerful approaches for association mapping in complex populations [both General Linear Model (GLM) and Mixed Linear Model (MLM)]. In the latest version, speed and memory efficiency have been important foci so that it can now handle millions of SNPs and conduct powerful tests with them much more rapidly than general purpose analysis packages.

TASSEL is an Open Source software, and can be found at <http://sourceforge.net/projects/tassel>

Please try to use informative subject lines for queries/comments! For bug reports please include as much information as possible including any error messages, the operating system, TASSEL version number, and whether you are using the stand-alone or web interface.

Rescueing data from a badd sequencing run  
By Nick Bell - 3 posts - 5 views 8:42 AM

Tassel 5 Requiring Java 8 next Release  
By Terry - 1 post - 0 views Aug 8

New! Tassel 5.0 Pipeline Document  
By Terry - 1 post - 4 views

# Resources for Tassel Users

- Available via [www.maizegenetics.net/tassel](http://www.maizegenetics.net/tassel)

The image shows a screenshot of a YouTube channel page for 'Panzea'. The channel name is 'Panzea' and it has 0 subscribers. The page displays three recent videos:

- 2. TASSEL menus** by Panzea, 1 day ago, 2 views. Video duration: 1:22.
- 3. TASSEL Data menu** by Panzea, 1 day ago, 3 views. Video duration: 3:53.
- 1. TASSEL installation and increasing heap size** by Panzea, 1 day ago, 3 views. Video duration: 3:39. Description: How to install TASSEL 5 on Mac OS and how to increase the heap size.

# Resources for Tassel Developers

## Wiki

Clone wiki Create page

Tassel 5 Source / Home

View History Edit

## TASSEL 5

### Developer Documentation

- [NetBeans Setup](#)
- [Guide for programming](#)
- [Tassel 5 Architecture](#)
- [Tassel 5 JavaDoc](#)
- [Creating Tassel Build from Source Code](#)
- [Tassel 5 Logging](#)
- [Executing Tassel 5 Unit Tests](#)
- [Tassel 5 Self-Describing Plugins](#)
- [Testing Tassel with JUnit](#)
- [Tassel and R Integration](#)

### Example Java Code

- [Reading/writing hapmap, manipulating genotype tables, writing to text file](#)

# 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** Dashboards ▾ Projects ▾ Issues ▾ Agile ▾ Create Search 🔍 ? 👤

## Kanban board

QUICK FILTERS: Only My Issues Recently Updated Board ▾

**4 Suggested**      **6 To Do**      **8 In Progress**      **42 Done**      Release...

- TAS-114** Filter SNPs not in LD with their neighbors
- TAS-116** Text output of a region of interest from HDF5 TOPM
- TAS-118** Discovery output every site with annotations
- TAS-334** Create TOPM Graphical Viewer
- TAS-386** improved output for ProductionSNPCallerPlugin
- TAS-417** Biology-based unit tests
- TAS-517** Code SetSNPQualityScores  
GBS pipeline v2
- TAS-520** Unit tests for DiscoverySNPCallerPlugin  
GBS pipeline v2
- TAS-702** Add pipeline parameters to remove DB table entries
- TAS-756** Create genome of first 20MB chr 9/10
- TAS-391** GBS pipeline v2
- TAS-571** SAMToGBSdbPlugin: Add filters for quality and identify quality  
GBS pipeline v2
- TAS-653** Investigate why SAMToDBPlugin is slow
- TAS-682** SAMToGBSdbPlugin junits failing
- TAS-696** SNPQuality table unique constraint
- TAS-734** GBS export of indels needs to fully define the ref and alt for VCF  
GBS pipeline v2
- TAS-507** Genome sequence class
- TAS-391 GBS pipeline v2
- TAS-564** Add Code to compare/process SNPs with conserved SNPs
- TAS-559** Organize GERP test data for CHR 9 & 10  
GBS pipeline v2
- TAS-404** One step TBT  
GBS pipeline v2
- TAS-518** Code ProductionSNPCallerPlugin  
GBS pipeline v2
- TAS-476** Design of GBSDData Interface  
GBS pipeline v2

# JIRA Issue

Tassel / TAS-768  
**Load Jason Wallace's GWAS hits**

Edit Comment Assign More ▾ Backlog Selected for Development Workflow ▾ Export ▾

### Details

Type:	Task	Status:	<b>DONE</b> <a href="#">(View Workflow)</a>
Priority:	Major	Resolution:	Done
Affects Version/s:	5	Fix Version/s:	None
Component/s:	<a href="#">Genomics</a>		
Labels:	None		
Epic Link:	<a href="#">GenomeAnnotations</a>		

### People

Assignee:	Jeff Glaubitz
Reporter:	Jeff Glaubitz
Votes:	0
Watchers:	<b>1</b> <a href="#">Stop watching this issue</a>

### 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.

### Dates

Created:	Yesterday
Updated:	Yesterday
Resolved:	Yesterday

### Activity

All **Comments** Work Log History Activity

There are no comments yet on this issue.

Comment

### Development

<a href="#">1 commit</a>	Latest Yesterday
<a href="#">Create branch</a>	

### Agile

[View on Board](#)



# Bitbucket/JIRA Integration

**JIRA** Dashboards ▾ Projects ▾ Issues ▾ Agile ▾ **Create** Search 🔍 ? 👤 ▾

**T** Tasse / **TAS-768**  
Load Jason Wallace's GWAS hits

Edit Comment Assign More ▾ Backlog Selected for Development Workflow ▾ Export ▾

### Details

Type:	Task	Status:	<b>DONE</b> <a>View Workflow</a>
Priority:	Major	Resolution:	Done
Affects Version/s:	5	Fix Version/s:	None
Component/s:	Genomics		
Labels:	None		
Epic Link:	GenomeAnnotations		

### 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 Glaubit
Reporter:	Jeff Glaubit
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

# Bitbucket/JIRA Integration

TAS-768: 1 unique commit



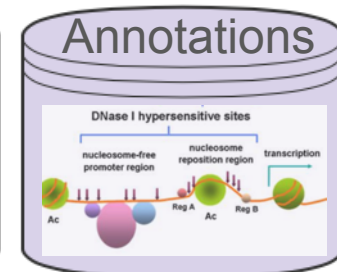
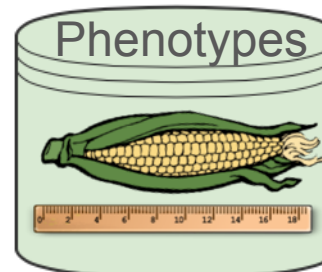
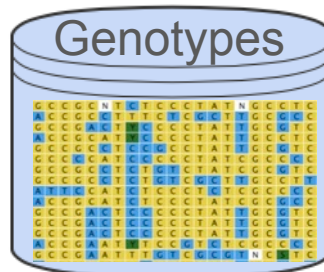
PrivateMaizegenetics

[Hide files](#)

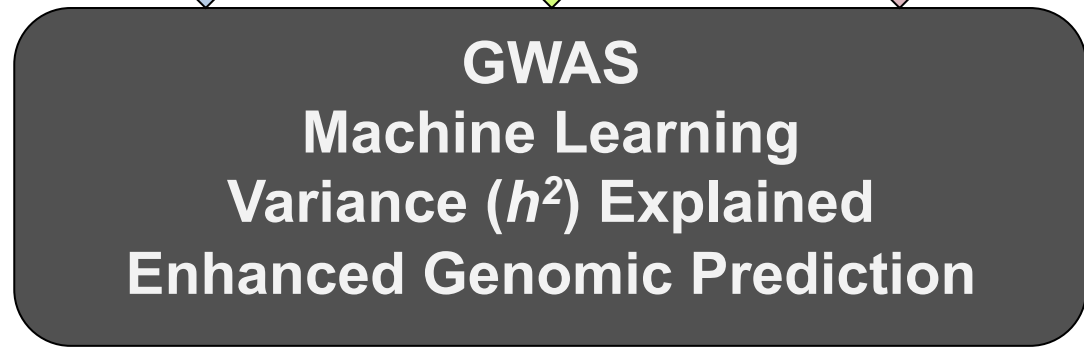
Author	Commit	Message	Date	Files
	<a href="#">2350a7a</a>	TAS-768 : Loaded Jason Wallace's GWAS hits into a gwas_hits table	Yesterday	<a href="#">2 files</a>
+235	-4	<code>src/net/maizegenetics/jeff/t5/GenomeAnnosDB.java</code>		
+19	-0	<code>src/net/maizegenetics/jeff/t5/maizeGenomeAnnos.sql</code>		

[Close](#)

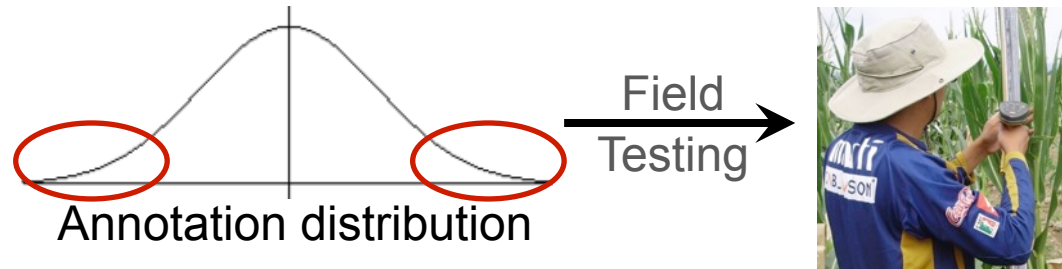
**Data:**



**Analysis:**



**Verification:**

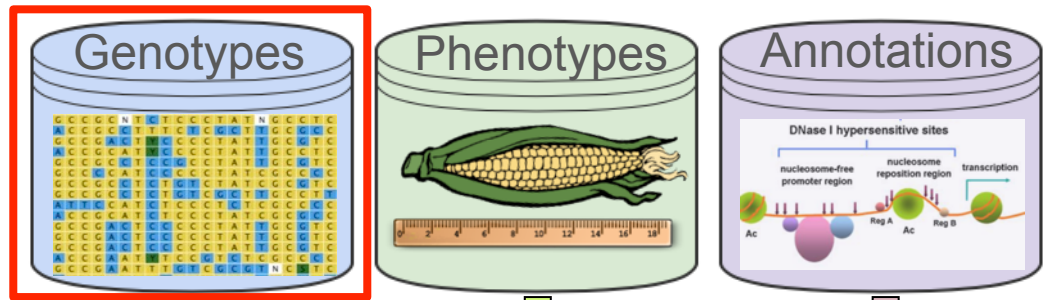


**Products:**

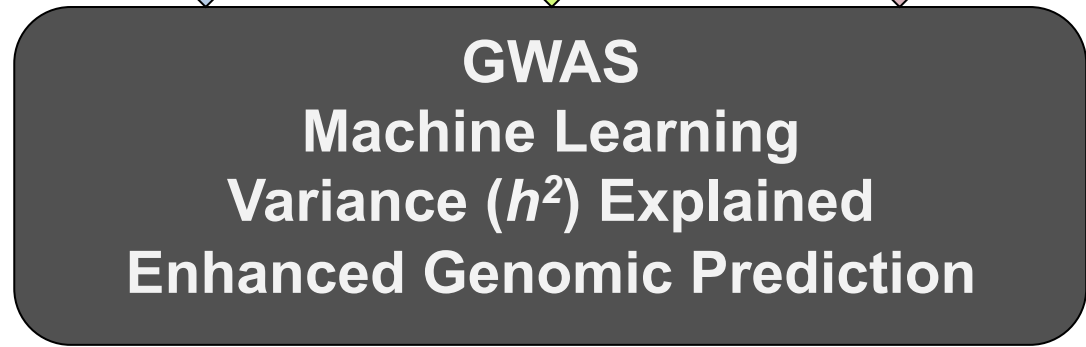


Personalized  
Medicine for  
Corn?

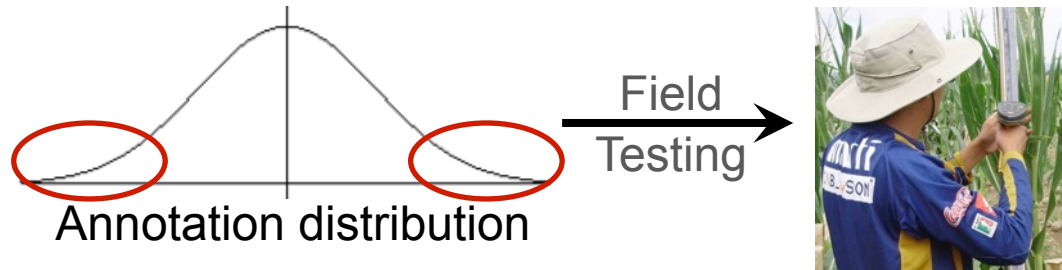
**Data:**



**Analysis:**



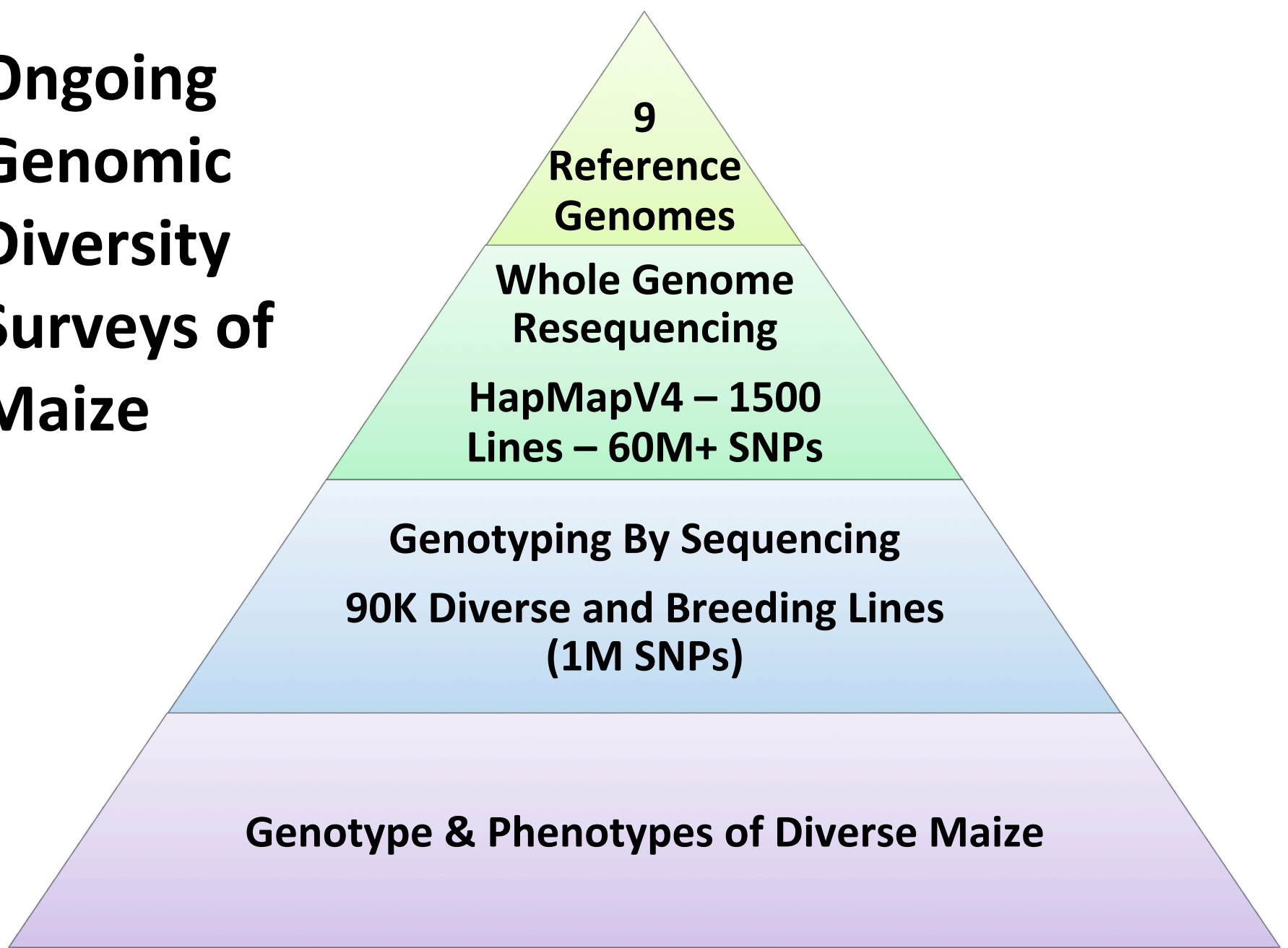
**Verification:**



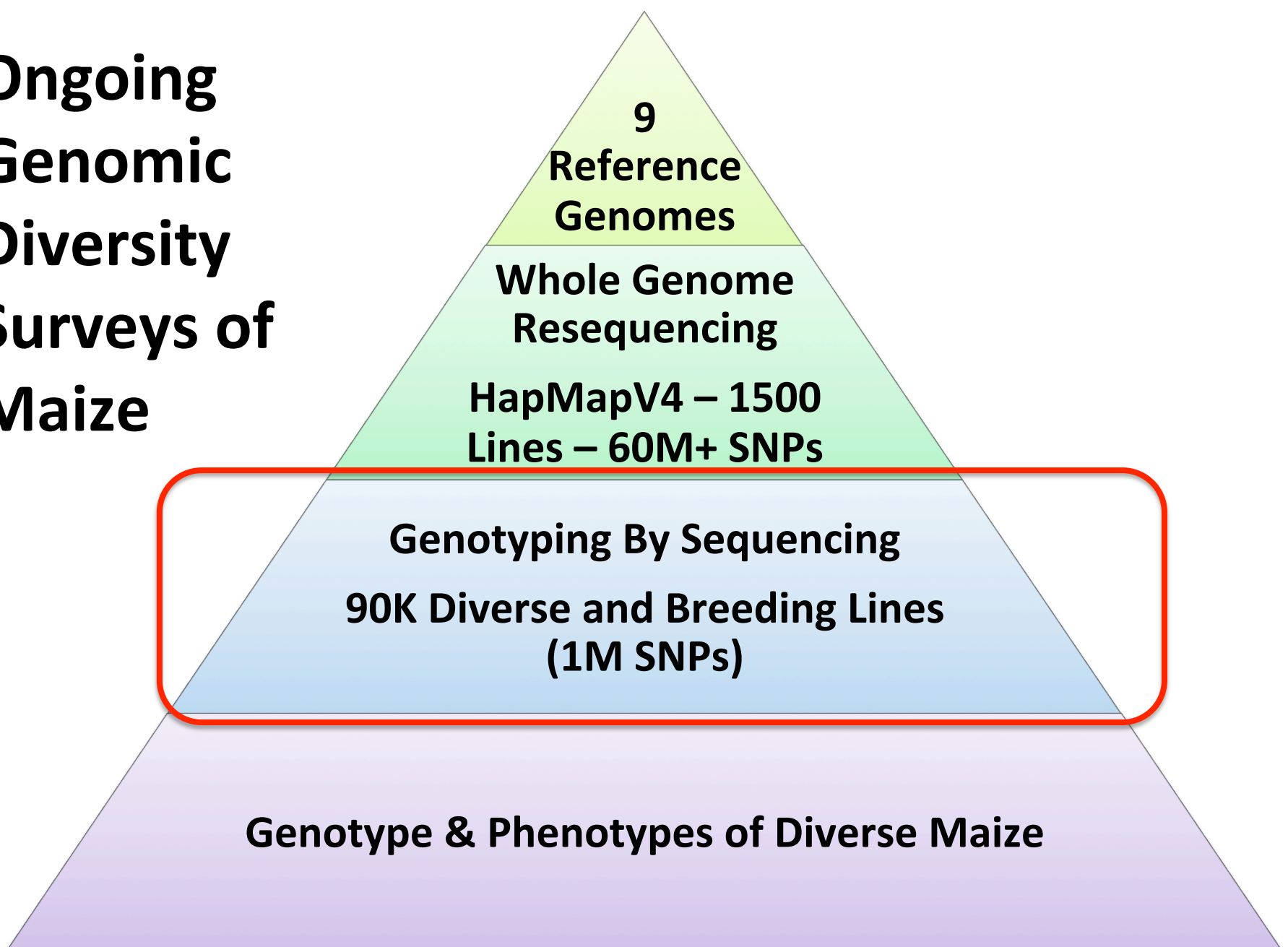
**Products:**



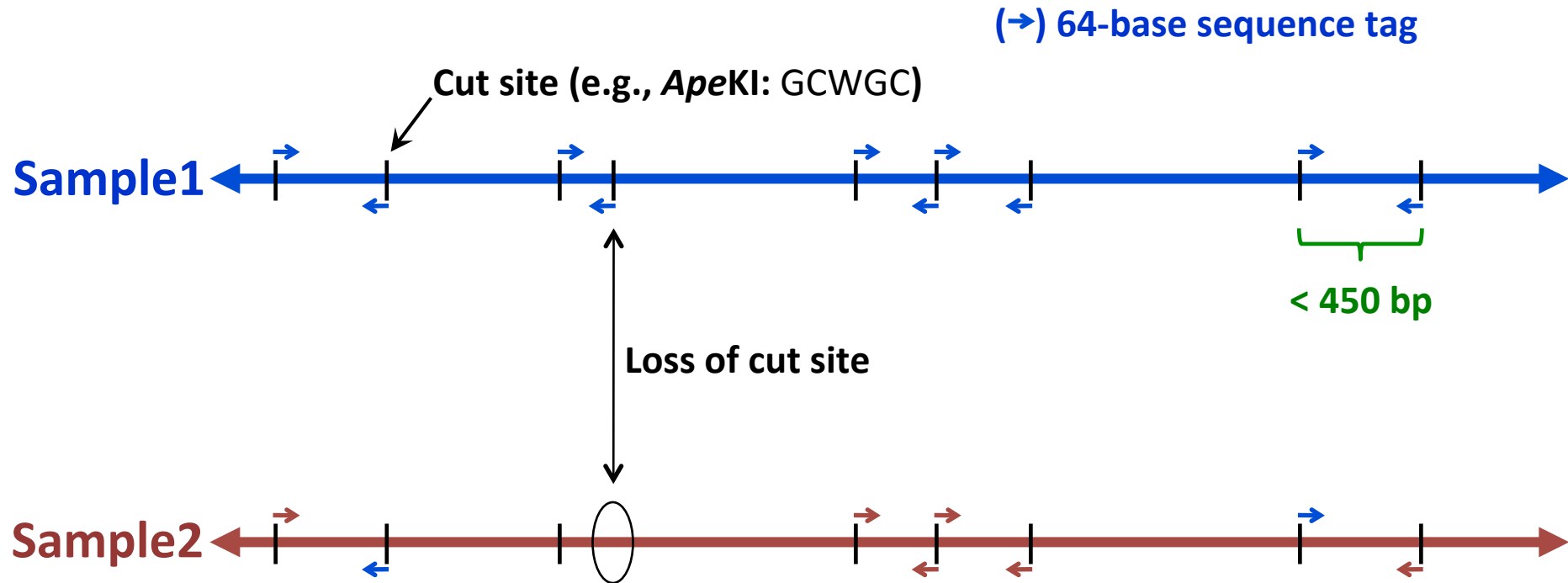
# Ongoing Genomic Diversity Surveys of Maize



# Ongoing Genomic Diversity Surveys of Maize



# What is genotyping-by-sequencing (GBS)?



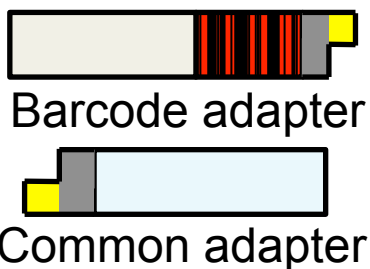
- Reduced representation approach inspired by Altshuler *et al.* (2000)
- Focuses NextGen sequencing power to ends of restriction fragments
- Scores both biallelic markers and presence/absence markers

# The GBS protocol is simple and robust

Elshire *et al.* 2011. PLoS ONE



Rob Elshire

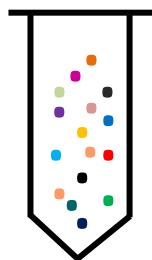


1. Plate DNA & adapter pair



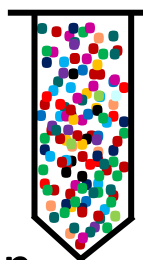
2. Digest DNA with RE  
(e.g. *ApeKI*)  
3. Ligate adapters

4. Pool DNAs

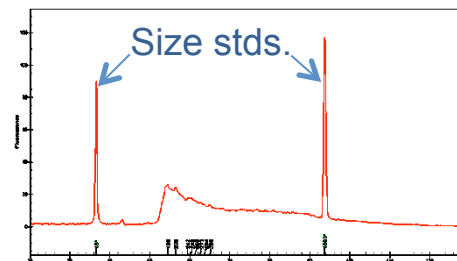


Primers

5. PCR  
6. Cleanup



7. Evaluate fragment sizes

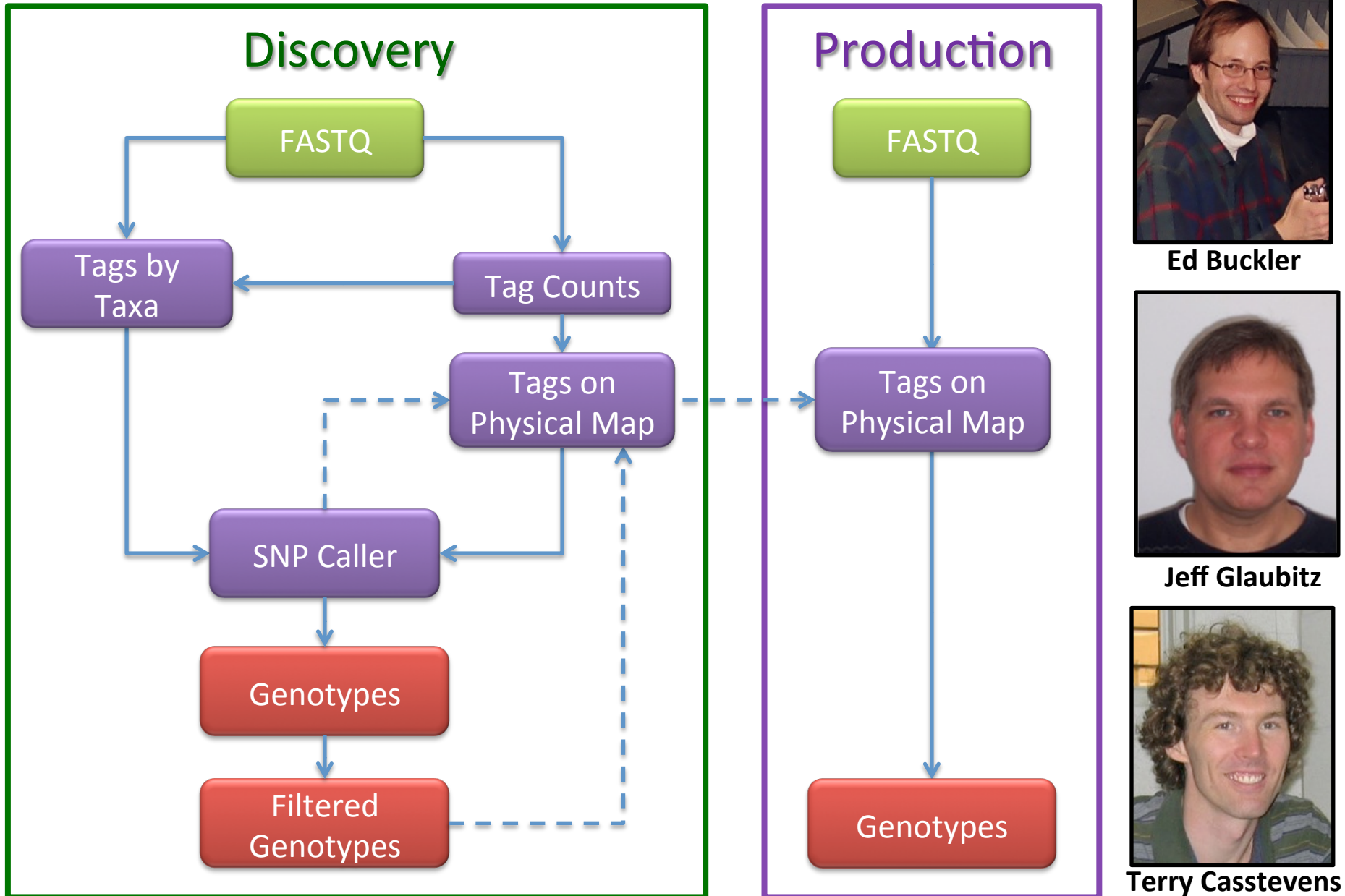


8. Run on single lane  
of Illumina flowcell

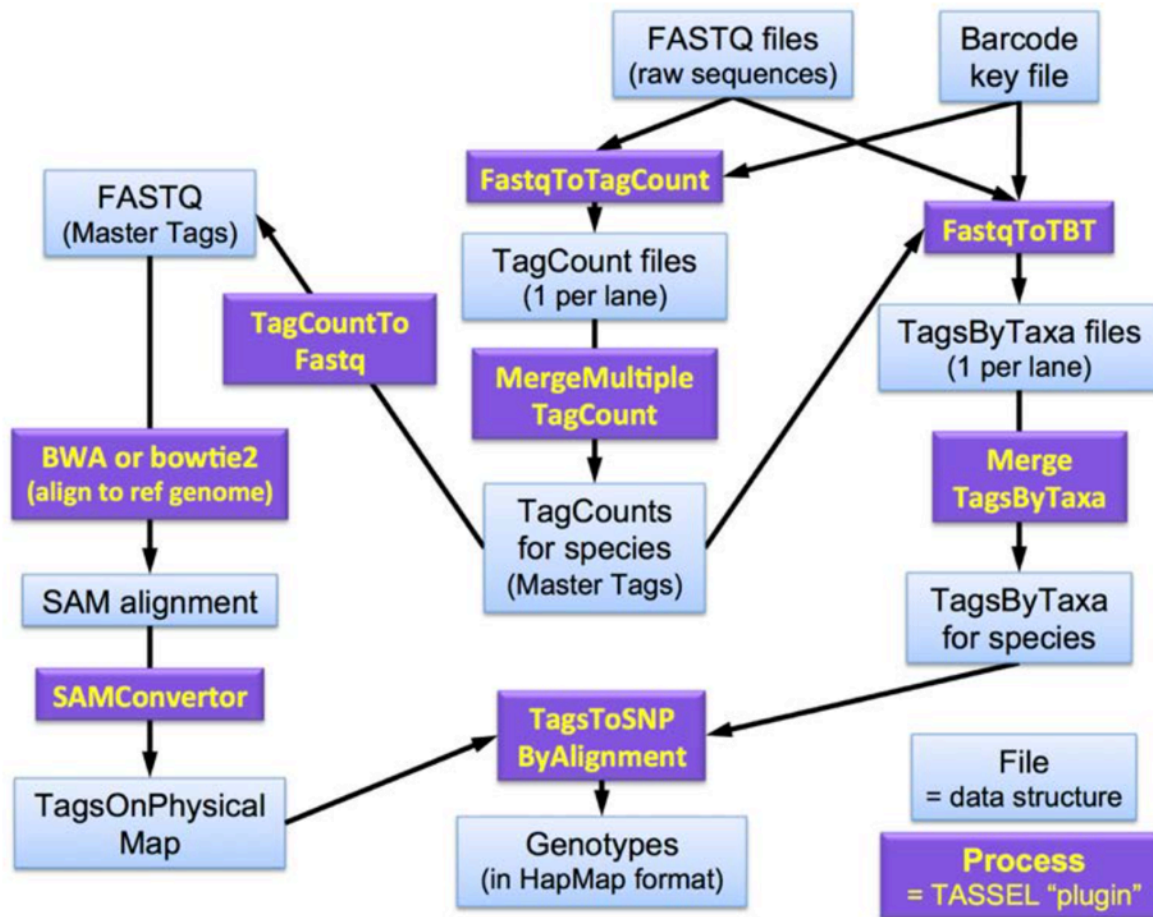
Sequence Reads in FASTQ format

# Reference-based GBS bioinformatics pipeline

Glaubitz *et al.* 2014 PLoS ONE

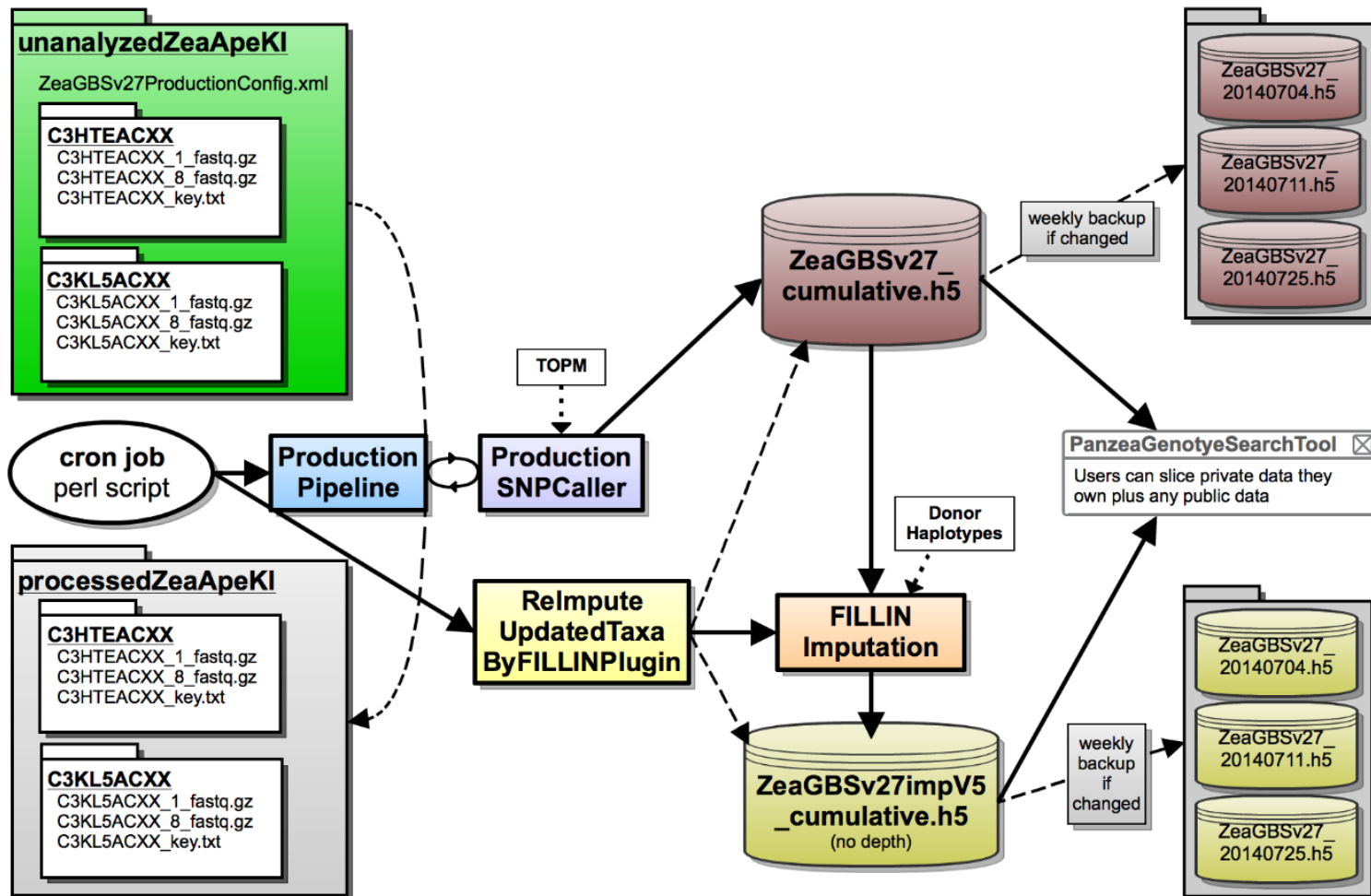


# TASSEL-GBSv1 Discovery Pipeline details



# 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



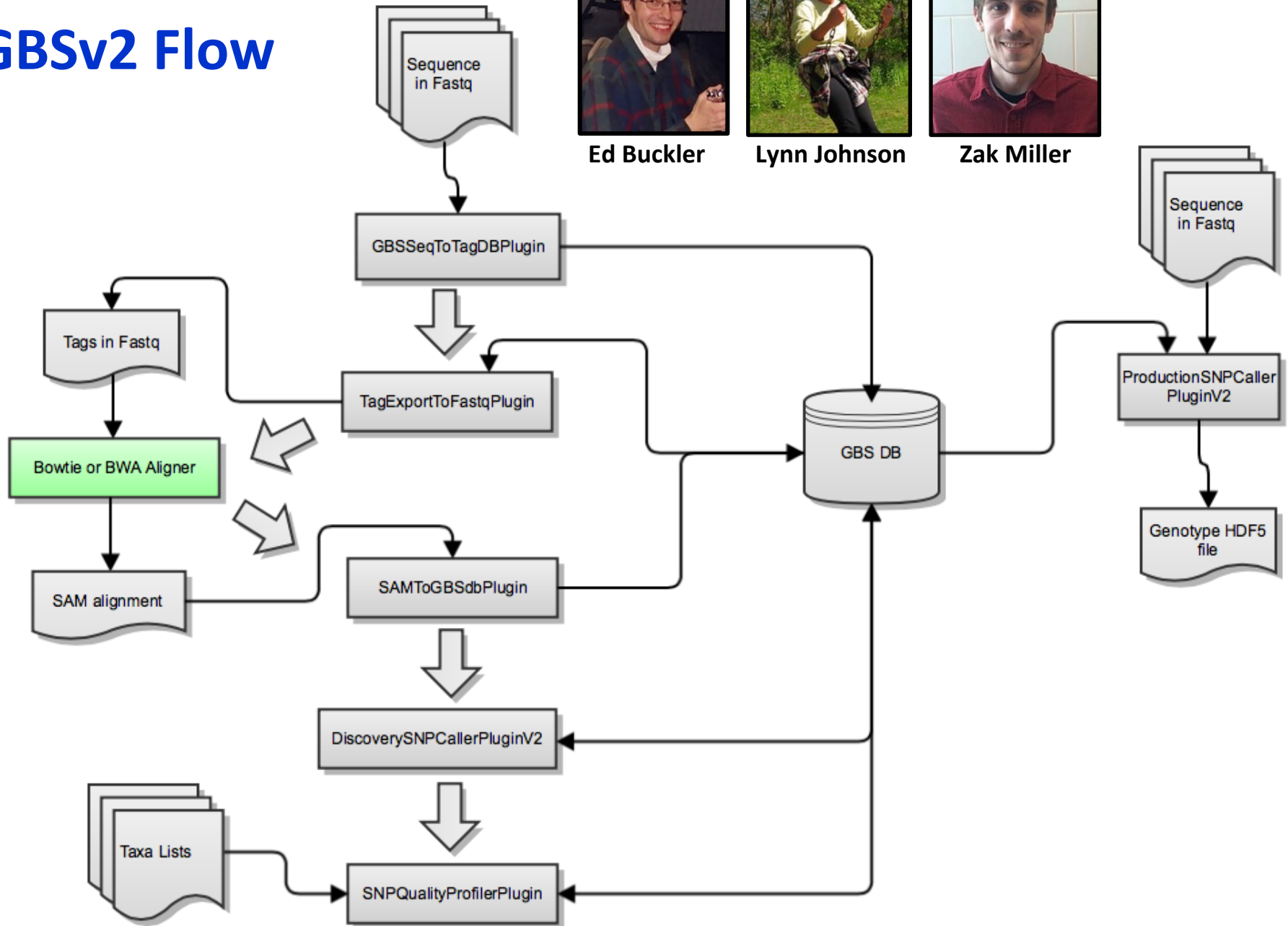
Ed Buckler



Lynn Johnson



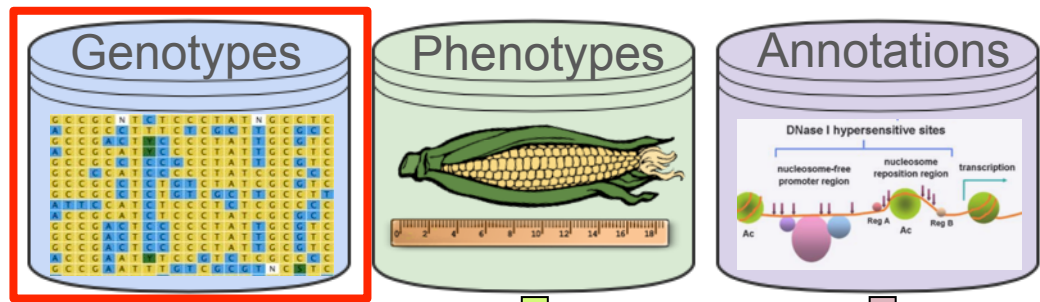
Zak Miller



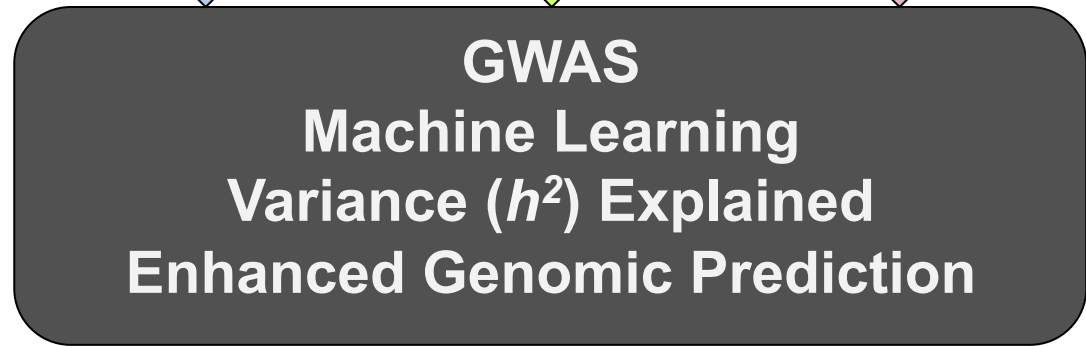
# Which SNPs to keep?

- 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

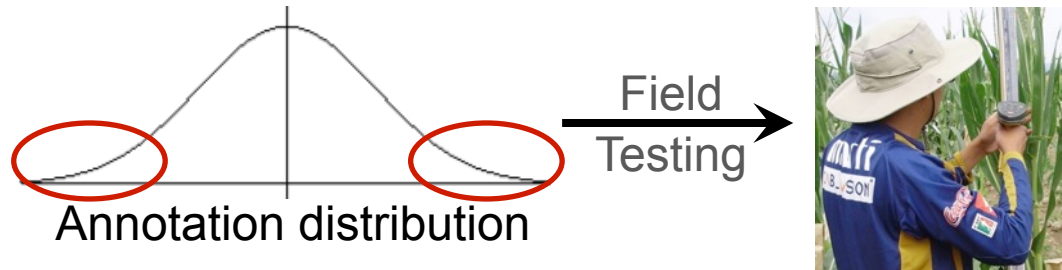
**Data:**



**Analysis:**



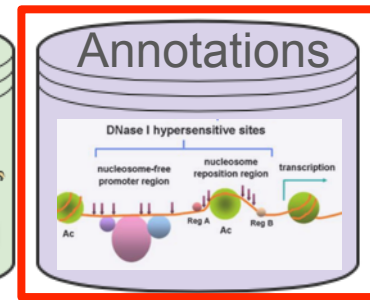
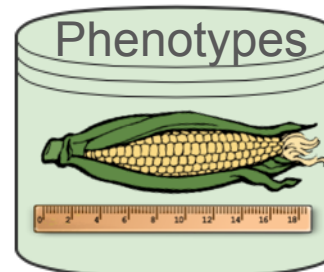
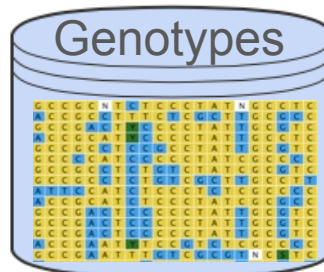
**Verification:**



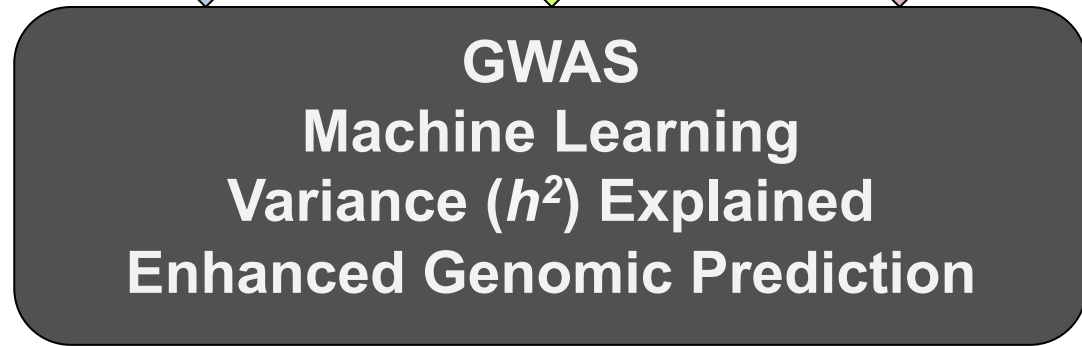
**Products:**



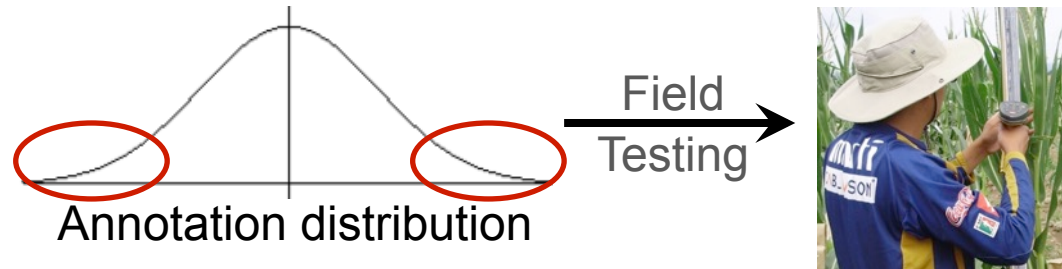
**Data:**



**Analysis:**



**Verification:**




**Products:**



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)