# COGENFITO: a composite genotype finder tool for optimizing isoline selection in maize breeding schemes

David Hessel [1,2,], Carolyn Lawrence [1,3,4], and Nick Lauter [1,2,3*]

[1] Interdepartmental Genetics Program, Iowa State University, Ames, IA, 50011, U.S.A.
[2] Department of Plant Pathology, Iowa State University, Ames, IA, 50011, U.S.A
[3] USDA-ARS Corn Insects and Crop Genetics Research, Ames, IA, 50011, U.S.A.
[4] Department of Genetics, Development and Cell Biology, Iowa State University, Ames, IA, 50011, U.S.A.

[*] Author to whom correspondence should be addressed: nick.lauter@ars.usda.gov

COGENFITO, the composite genotype finder tool, is an interactive browser that increases accessibility and utility of public genotype datasets stored at MaizeGDB. COGENFITO efficiently sorts and sifts through genetic marker data to identify lines with user-defined combinations of alleles. Currently, COGENFITO interrogates genotype data for 26 RIL populations (IBM RILs + NAM RILs) for which seed stocks are freely available from the Maize Genetics Cooperation – Stock Center (MGCSC). Researchers need only specify the mapping population and desired genotypes at markers of interest in order to get a result. Browsing the genotype data associated with particular local areas of maps is facilitated by a "show centiMorgan range" function. A color-coded graphical display of genotypes facilitates visual and intuitive selection of lines, which each have specific links to the MGCSC. COGENFITO is applicable to a wide range of research pursuits, including QTL cloning, dissection of epistatic interactions, and control of genetic background in selection and improvement schemes.

## Introduction

Genetic marker data are less expensive and more efficient to collect than ever before, and the rate of improvement continues to increase (Prigge et al., 2009). Together with improved protocols for creating doubled haploids (Longin et al., 2006) and for kernel chipping, both of which allow genotypes to be known prior to planting, these gains are changing not only what is possible, but also what is cost-effective for commercial breeding. The vast increase in amount of genotype data collected both in the maize research and breeding sectors has created the need for better data management tools for display and interrogation. In the research sector, this applies primarily to genotyped sets of Recombinant Inbred Lines (RILs) or doubled haploids (DHs) for which seeds are available from the MGCSC or North Central Regional Plant Introduction Station. Until now, researchers had to download and format genotype datasets in order to use the data for making decisions. Since datasets are frequently updated, such snapshots become outdated quite quickly, requiring additional retrieval and formatting . Users also needed to invent their own strategies for sorting and sifting through data, usually requiring hand annotation of the markers of interest by viewing relevant map displays at MaizeGDB. Here we introduce COGENFITO and demonstrate how it can meet these needs. Example-based tutorials for using the tools and several use case scenarios for experimental breeding are provided. Finally, we discuss future directions for extending the COGENFITO tool set to make it more directly suited to the needs of breeders.

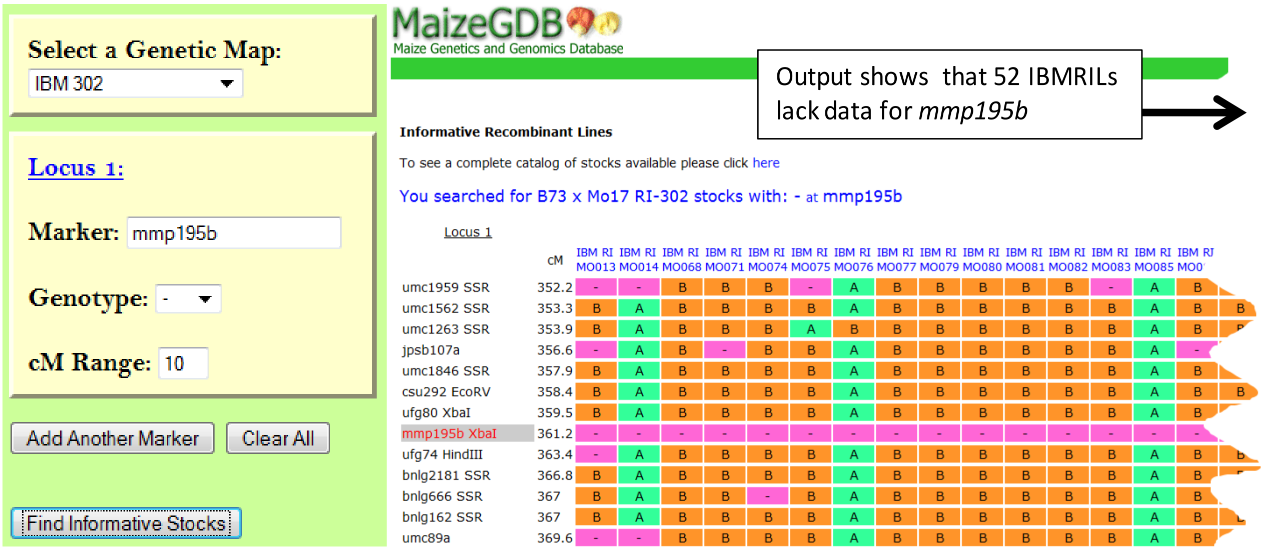
## Results

COGENFITO has been implemented within the context of MaizeGDB (Lawrence et al., 2007) and is currently available for use in browsing genotype data from 26 different RIL populations of maize (http://maizegdb.org/Cogenfito). The populations include the 25 Nested Association Mapping (NAM) RIL sets (McMullen et al., 2009) and the Intermated B73 x Mo17 (IBM) RILs (Lee et al., 2002). Together, these comprise nearly 5,000 lines that currently have more than 1,100 marker data points each.

*Data types and usage rules*: The data for the IBM RILs are mainly simple sequence repeat (SSR) and insertion/deletion polymorphism (indel) co-dominant markers that produce simple sequence length polymorphisms (SSLPs) detected by gel or capillary electrophoresis. However, cleaved amplified polymorphic site (CAPS), restriction fragment length polymorphism (RFLP) and single nucleotide polymorphism (SNP) data are also abundant among the more than 9,000 markers that have been placed on the IBMRILs (Sharopova et al., 2002; Fu et al., 2006). By contrast, the NAM RIL data sets are almost exclusively SNP-based (McMullen et al., 2009). In order to handle all of these data types in a unified way, the data are stored in MaizeGDB with alternate homozygous classes specified as A and B and missing data denoted by a dash. By definition, the A genotype represents plants that are homozygous for the allele donated by the female plant in the original cross. Currently, this organization does not present a logistical challenge for researchers using COGENFITO because all 26 populations have B73 in common as the female parent.

*Tutorial for browsing genotype data:* In the era of genomics, we often utilize data that we know little about, or have never even seen. Oftentimes when an anomalous result is observed, an important first step is to check the data underpinning it. For example, while mapping maturity traits in the IBM RILs, Lauter and colleagues (2008) noticed a deep cleft in the QTL peak for the *vegetative to generative transition1* locus in chromosome bin 8.05 (Salvi et al., 2007). They suspected that the cleft in the QTL likelihood curve was due to poor quality genotype data at *mmp195b*, an RFLP marker for which 6 banding patterns were assigned genetic positions. BLAST (Altschul et al., 1990) searches of the expressed sequence tag (EST) associated with the RFLP marker revealed only a very weak match to the physical region near *vgt1*, suggesting that both incorrect genotype calls and missing data may have been to blame (Lauter et al., 2008). COGENFITO makes examination of the extent of missing data for the *mmp195b* marker in the IBMRILs easy by allowing a researcher to specify the population and marker, as well as a cM range surrounding it (Figure 1).

**Figure 1.** IBMRILs with missing data at *mmp195b* plus data from other markers within 10 cM



Select a Genetic Map: IBM 302 ▼

Locus 1:

Marker: mmp195b

Genotype: - ▼

cM Range: 10

Output shows that 52 IBMRILs lack data for *mmp195b*

Informative Recombinant Lines

To see a complete catalog of stocks available please click here

You searched for B73 x Mo17 RI-302 stocks with: - at mmp195b

Locus 1

| | cM | IBM RI MO013 | IBM RI MO014 | IBM RI MO068 | IBM RI MO071 | IBM RI MO074 | IBM RI MO075 | IBM RI MO076 | IBM RI MO077 | IBM RI MO079 | IBM RI MO080 | IBM RI MO081 | IBM RI MO082 | IBM RI MO083 | IBM RI MO085 | IBM RJ MO0 | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| umc1959 SSR | 352.2 | - | - | B | B | B | - | A | B | B | B | B | B | - | A | B | |
| umc1562 SSR | 353.3 | B | A | B | B | B | B | A | B | B | B | B | B | B | A | B | B |
| umc1263 SSR | 353.9 | B | A | B | B | B | A | B | B | B | B | B | B | B | A | B | B |
| jpsb107a | 356.6 | - | A | B | - | B | B | A | B | B | B | B | B | B | A | - | |
| umc1846 SSR | 357.9 | B | A | B | B | B | B | A | B | B | B | B | B | B | A | B | |
| csu292 EcoRV | 358.4 | B | A | B | B | B | B | A | B | B | B | B | B | B | A | B | B |
| ufg80 XbaI | 359.5 | B | A | B | B | B | B | A | B | B | B | B | B | B | A | B | |
| mmp195b XbaI | 361.2 | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | |
| ufg74 HindIII | 363.4 | - | A | B | B | B | B | A | B | B | B | B | B | B | A | B | B |
| bnlg2181 SSR | 366.8 | B | A | B | B | B | B | A | B | B | B | B | B | B | A | B | |
| bnlg666 SSR | 367 | B | A | B | B | - | B | A | B | B | B | B | B | B | A | B | |
| bnlg162 SSR | 367 | B | A | B | B | B | B | A | B | B | B | B | B | B | A | B | B |
| umc89a | 369.6 | - | - | B | B | B | B | A | B | B | B | B | B | B | A | B | |

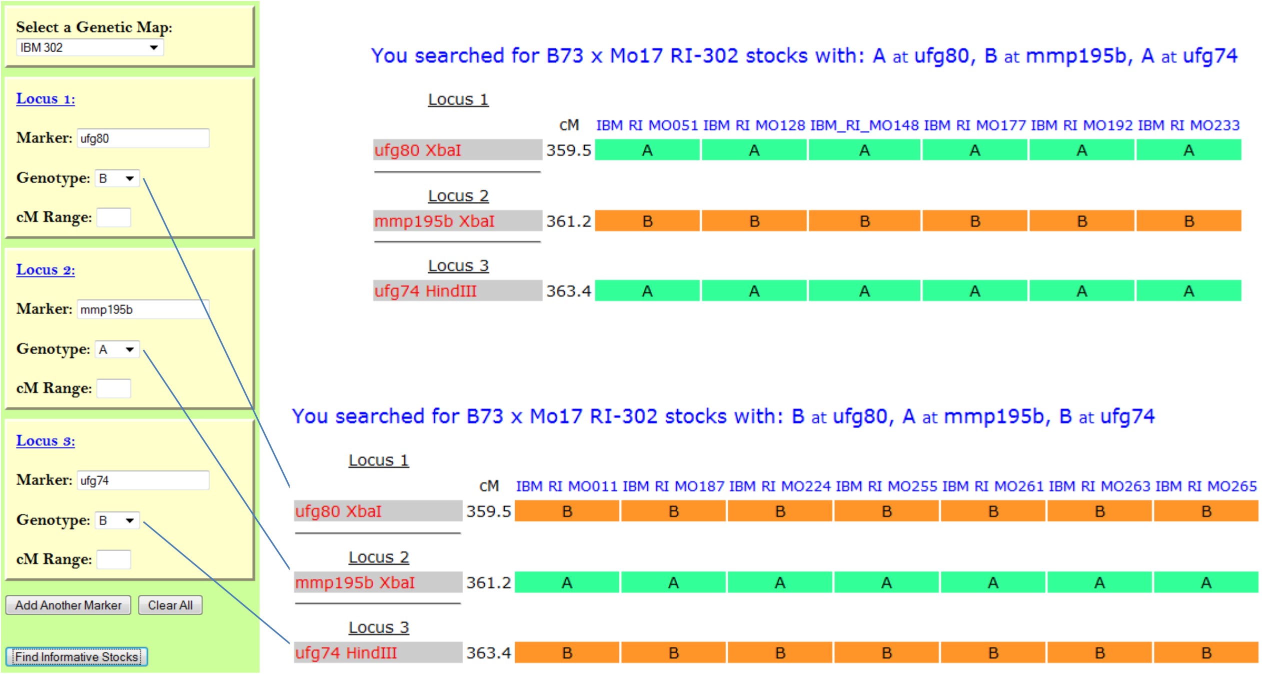
Input page showing Map, Marker, Genotype and cM range specified

Relevant portion of the output page showing a label summarizing the query and the table of genotype data in specified range for matching lines.

To consider whether or not *mmp195b* may have been incorrectly forced into this position on the map, one could investigate the numbers of double crossovers required by positioning it between *ufg80* and *ufg74* that are not supported by any other adjacent markers. To do this, use the "Add Another Marker" feature to specifying the appropriate strings of genotypes (Figure 2). This type of usage is ideal for investigating genetic evidence for marker order in a map as well; it is common to observe a physical order of X-Y-Z for three markers in the physical map, but an X-Z-Y order in the genetic map. Investigating the data to see how strong the genetic evidence is for the reported order can be a reasonable first step and is facilitated by COGENFITO.

<u>*Tutorial for searching for isolines with specific multilocus genotypes:*</u> The examples provided in Figures 1 and 2 have respectively demonstrated the use of the “cM range” and “Add Another Marker” utilities of COGENFITO. We have found that it is often useful to combine these two features when performing a search for lines with a particular composite genotype. The former is used so that a large block of marker data surrounding a specific position of interest is displayed, while the latter is used to apply filtering within the primary locus of interest as well as at background markers whose genotype must also be controlled as part of the particular breeding scheme being employed.

**Figure 2.** Thirteen double crossovers result from the original placement of *mmp195b*



Select a Genetic Map: IBM 302

Locus 1:

Marker: ufg80

Genotype: B

cM Range:

Locus 2:

Marker: mmp195b

Genotype: A

cM Range:

Locus 3:

Marker: ufg74

Genotype: B

cM Range:

You searched for B73 x Mo17 RI-302 stocks with: A at ufg80, B at mmp195b, A at ufg74

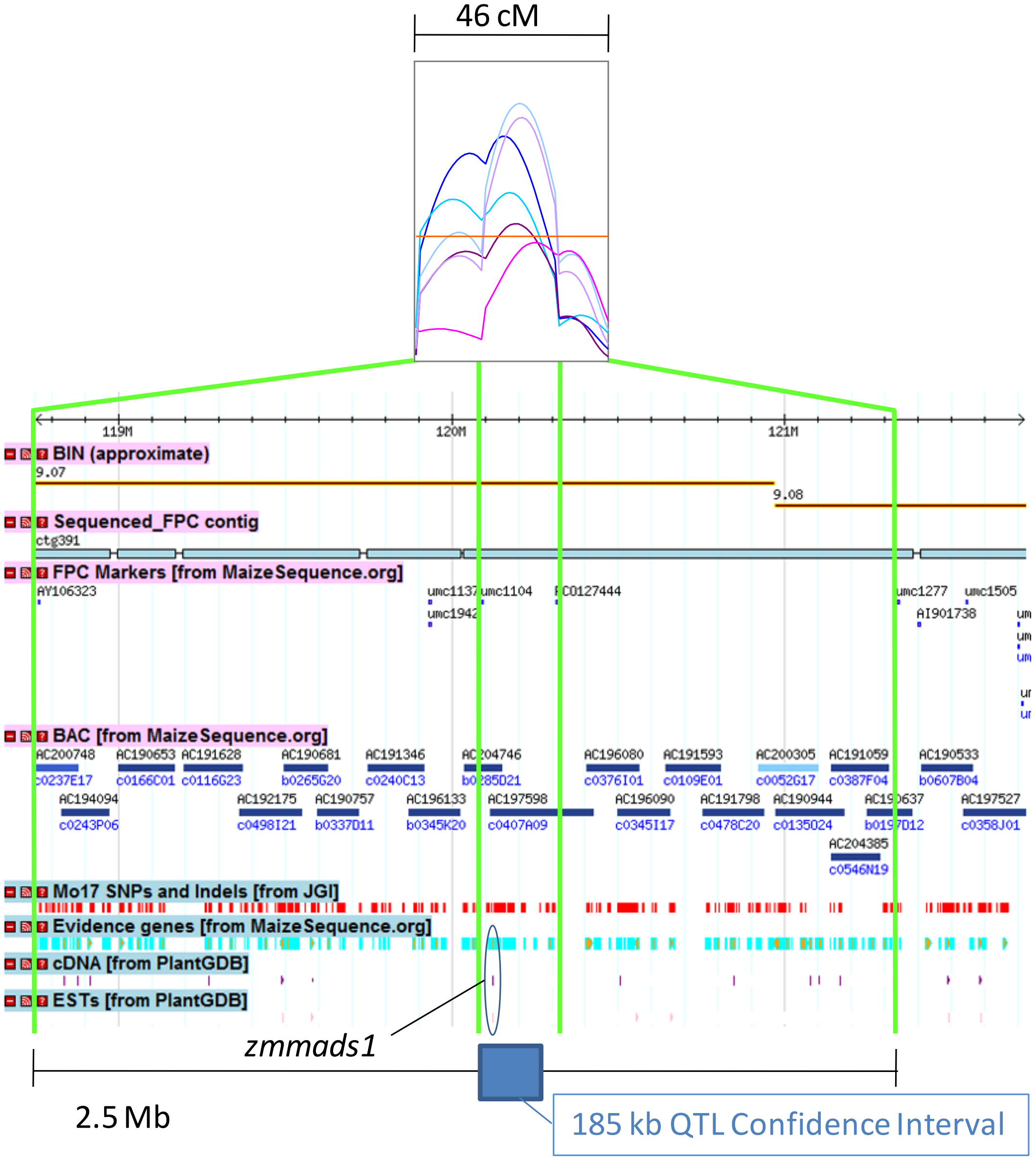
| | cM | IBM RI MO051 | IBM RI MO128 | IBM_RI_MO148 | IBM RI MO177 | IBM RI MO192 | IBM RI MO233 |
|---|---|---|---|---|---|---|---|
| Locus 1 | | | | | | | |
| ufg80 XbaI | 359.5 | A | A | A | A | A | A |
| Locus 2 | | | | | | | |
| mmp195b XbaI | 361.2 | B | B | B | B | B | B |
| Locus 3 | | | | | | | |
| ufg74 HindIII | 363.4 | A | A | A | A | A | A |

You searched for B73 x Mo17 RI-302 stocks with: B at ufg80, A at mmp195b, B at ufg74

| | cM | IBM RI MO011 | IBM RI MO187 | IBM RI MO224 | IBM RI MO255 | IBM RI MO261 | IBM RI MO263 | IBM RI MO265 |
|---|---|---|---|---|---|---|---|---|
| Locus 1 | | | | | | | | |
| ufg80 XbaI | 359.5 | B | B | B | B | B | B | B |
| Locus 2 | | | | | | | | |
| mmp195b XbaI | 361.2 | A | A | A | A | A | A | A |
| Locus 3 | | | | | | | | |
| ufg74 HindIII | 363.4 | B | B | B | B | B | B | B |

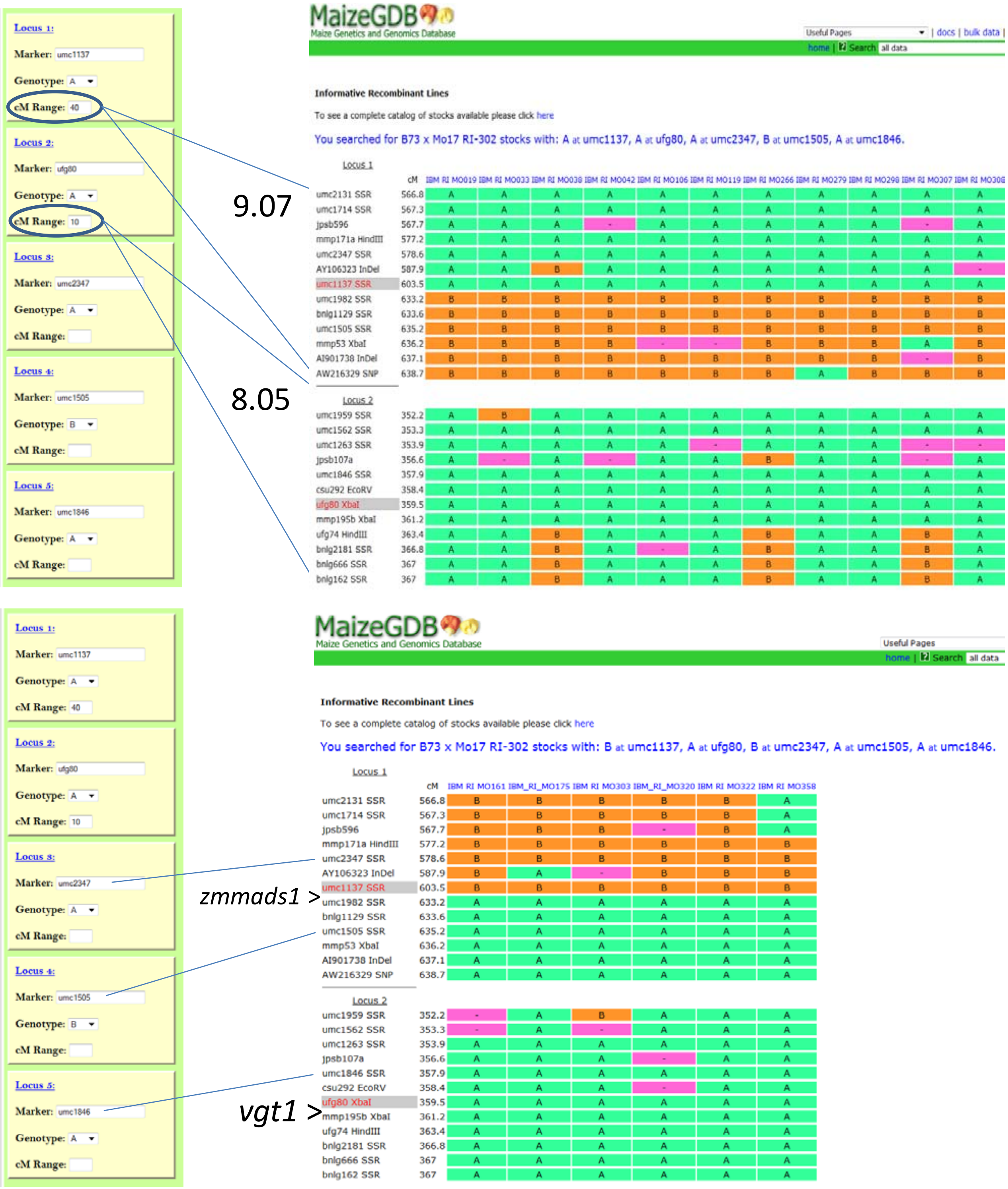
Consider the example pictured in Figure 3, where Lauter and colleagues (2008) identified a maize maturity QTL in bin 9.07 that localizes to a narrow physical interval with a high degree of recombination that is poorly resolved by markers. The MADS box gene, *zmmads1* is one of the few computationally predicted or evidence-based (expressed) genes that physically reside within the boundaries of a stringently defined 95% confidence interval (Lauter et al., 2008; Crossett et al., 2010). Thus, genetically interrogating the lines harboring recombination events adjacent to *zmmads1* is an obvious next step. Adding genetic markers to this region is facilitated by the SNP annotation track in the MaizeGDB BAC-Based Genome Browser (Sen et al.). However, due to the complexity of the genetic regulation of maturity traits, the effects of the 9.07 QTL are not Mendelian and thus require measurement in segregating progeny in order to be rigorously associated with any particular genetic polymorphism delineated by marker data.

**Figure 3**. Association of physical map with maturity QTL in bin 9.07



In this example, the genetic effects on maturity of *vgt* in bin 8.05 are so strong that they must be considered when making crosses to create progeny groups for testing the effects of the 9.07 QTL on maturity. The COGENFITO settings and outcomes of the search for recombinants within the genetic interval that harbors *zmmads1* while controlling the *vgt1* genotype are shown in Figure 4. More than a handful of A to B (upper panel) and B to A (lower panel) recombinants were found between umc1137 and umc1982. The upper panel is annotated to show how the use of the "cM range" feature keeps the data in large blocks. The lower panel is annotated to show how additional markers within bins 9.07 and 8.05 were respectively used to filter for recombinants and rigorously control the genotype at *vgt1*. For simplicity, several lines of output are omitted; see Figure 2 for comparison.

**Figure 4.** Identifying recombinants in bin 9.07 in lines with a high allele for *vgt1*


MaizeGDB
Maize Genetics and Genomics Database
Useful Pages | docs | bulk data |
home | Search all data
9.07
8.05

Locus 1:

Marker: umc1137

Genotype: A

cM Range: 40

Locus 2:

Marker: ufg80

Genotype: A

cM Range: 10

Locus 3:

Marker: umc2347

Genotype: A

cM Range:

Locus 4:

Marker: umc1505

Genotype: B

cM Range:

Locus 5:

Marker: umc1846

Genotype: A

cM Range:

**Informative Recombinant Lines**

To see a complete catalog of stocks available please click here

You searched for B73 x Mo17 RI-302 stocks with: A at umc1137, A at ufg80, A at umc2347, B at umc1505, A at umc1846.

Locus 1

| | cM | IBM RI MO019 | IBM RI MO033 | IBM RI MO038 | IBM RI MO042 | IBM RI MO106 | IBM RI MO119 | IBM RI MO266 | IBM RI MO279 | IBM RI MO298 | IBM RI MO307 | IBM RI MO308 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| umc2131 SSR | 566.8 | A | A | A | A | A | A | A | A | A | A | A |
| umc1714 SSR | 567.3 | A | A | A | A | A | A | A | A | A | A | A |
| jpsb596 | 567.7 | A | A | A | - | A | A | A | A | A | - | A |
| mmp171a HindIII | 577.2 | A | A | A | A | A | A | A | A | A | A | A |
| umc2347 SSR | 578.6 | A | A | A | A | A | A | A | A | A | A | A |
| AY106323 InDel | 587.9 | A | A | B | A | A | A | A | A | A | A | - |
| umc1137 SSR | 603.5 | A | A | A | A | A | A | A | A | A | A | A |
| umc1982 SSR | 633.2 | B | B | B | B | B | B | B | B | B | B | B |
| bnlg1129 SSR | 633.6 | B | B | B | B | B | B | B | B | B | B | B |
| umc1505 SSR | 635.2 | B | B | B | B | B | B | B | B | B | B | B |
| mmp53 XbaI | 636.2 | B | B | B | B | - | - | B | B | B | A | B |
| AI901738 InDel | 637.1 | B | B | B | B | B | B | B | B | B | - | B |
| AW216329 SNP | 638.7 | B | B | B | B | B | B | B | A | B | B | B |
| Locus 2 | | | | | | | | | | | | |
| umc1959 SSR | 352.2 | A | B | A | A | A | A | A | A | A | A | A |
| umc1562 SSR | 353.3 | A | A | A | A | A | A | A | A | A | A | A |
| umc1263 SSR | 353.9 | A | A | A | A | A | - | A | A | A | - | - |
| jpsb107a | 356.6 | A | - | A | - | A | A | B | A | A | - | A |
| umc1846 SSR | 357.9 | A | A | A | A | A | A | A | A | A | A | A |
| csu292 EcoRV | 358.4 | A | A | A | A | A | A | A | A | A | A | A |
| ufg80 XbaI | 359.5 | A | A | A | A | A | A | A | A | A | A | A |
| mmp195b XbaI | 361.2 | A | A | A | A | A | A | A | A | A | A | A |
| ufg74 HindIII | 363.4 | A | A | B | A | A | A | B | A | A | B | A |
| bnlg2181 SSR | 366.8 | A | A | B | A | - | A | B | A | A | B | A |
| bnlg666 SSR | 367 | A | A | B | A | A | A | B | A | A | B | A |
| bnlg162 SSR | 367 | A | A | B | A | A | A | B | A | A | B | A |

Locus 1:

Marker: umc1137

Genotype: A

cM Range: 40

Locus 2:

Marker: ufg80

Genotype: A

cM Range: 10

Locus 3:

Marker: umc2347

Genotype: A

cM Range:

Locus 4:

Marker: umc1505

Genotype: B

cM Range:

Locus 5:

Marker: umc1846

Genotype: A

cM Range:

**Informative Recombinant Lines**

To see a complete catalog of stocks available please click here

You searched for B73 x Mo17 RI-302 stocks with: B at umc1137, A at ufg80, B at umc2347, A at umc1505, A at umc1846.

Locus 1

| | cM | IBM RI MO161 | IBM_RI_MO175 | IBM RI MO303 | IBM_RI_MO320 | IBM RI MO322 | IBM RI MO358 |
|---|---|---|---|---|---|---|---|
| umc2131 SSR | 566.8 | B | B | B | B | B | A |
| umc1714 SSR | 567.3 | B | B | B | B | B | A |
| jpsb596 | 567.7 | B | B | B | - | B | A |
| mmp171a HindIII | 577.2 | B | B | B | B | B | B |
| umc2347 SSR | 578.6 | B | B | B | B | B | B |
| AY106323 InDel | 587.9 | B | A | - | B | B | B |
| umc1137 SSR | 603.5 | B | B | B | B | B | B |
| umc1982 SSR | 633.2 | A | A | A | A | A | A |
| bnlg1129 SSR | 633.6 | A | A | A | A | A | A |
| umc1505 SSR | 635.2 | A | A | A | A | A | A |
| mmp53 XbaI | 636.2 | A | A | A | A | A | A |
| AI901738 InDel | 637.1 | A | A | A | A | A | A |
| AW216329 SNP | 638.7 | A | A | A | A | A | A |
| Locus 2 | | | | | | | |
| umc1959 SSR | 352.2 | - | A | B | A | A | A |
| umc1562 SSR | 353.3 | - | A | - | A | A | A |
| umc1263 SSR | 353.9 | A | A | A | A | A | A |
| jpsb107a | 356.6 | A | A | A | - | A | A |
| umc1846 SSR | 357.9 | A | A | A | A | A | A |
| csu292 EcoRV | 358.4 | A | A | A | - | A | A |
| ufg80 XbaI | 359.5 | A | A | A | A | A | A |
| mmp195b XbaI | 361.2 | A | A | A | A | A | A |
| ufg74 HindIII | 363.4 | A | A | A | A | A | A |
| bnlg2181 SSR | 366.8 | A | A | A | A | A | A |
| bnlg666 SSR | 367 | A | A | A | A | A | A |
| bnlg162 SSR | 367 | A | A | A | A | A | A |

*Additional utilities of COGENFITO:* Oftentimes, QTL studies are restricted by the number of markers surveyed across the genome. When significant associations are identified using these smaller panels of markers, the confidence intervals for likely QTL position are often quite large yet strictly defined to a chromosomal region. COGENFITO can be deployed here to rapidly expedite the selection of subsequent markers for QTL finemapping. Frequently only a few carefully selected markers are required to drastically reduce the region of a QTL likelihood peak (Lauter et al., 2008).

In addition to mapping projects, several research pursuits depend on knowing the relative recombination frequency within a genomic region. These include, but are not limited to studies on transgenesis, silencing of heterochromatin, gene silencing, and molecular evolution. These all depend on measuring or estimating the amount of plasticity occurring in a specified genomic region. Chromosome-wide queries using COGENFITO can assess the relative recombination that has experimentally occurred across large genomic blocks. This is also a good way to identify which mapping population would best be suited for a particular QTL study if chromosome resolution of the QTL exists.

*Known caveats for use of COGENFITO:* There are several issues that may arise during use of COGENFITO. One of them is that specification of marker names is not always easy, as marker names often have versioning inconsistencies. To minimize problems associated with such discrepancies, we have made the entry of the marker names case-insensitive and have deployed an auto-completion mechanism that overcomes decimal-based versioning. If a marker name is not recognized or that marker was not on the genetic map for the population specified, the following error message will be displayed: "Marker is not found on the specified genetic map. Please refine your search parameters." Please note that if you have specified multiple markers, the error message does not indicate which marker is not found. For this reason, we recommend incrementally building complex searches that require individual specifications at many loci. When use of a particular marker is important and it is not recognized by COGENFITO, we recommend that you use the advanced map search tool to check both the map occurrence and the syntax for the marker of interest (http://www.maizegdb.org/map.php).

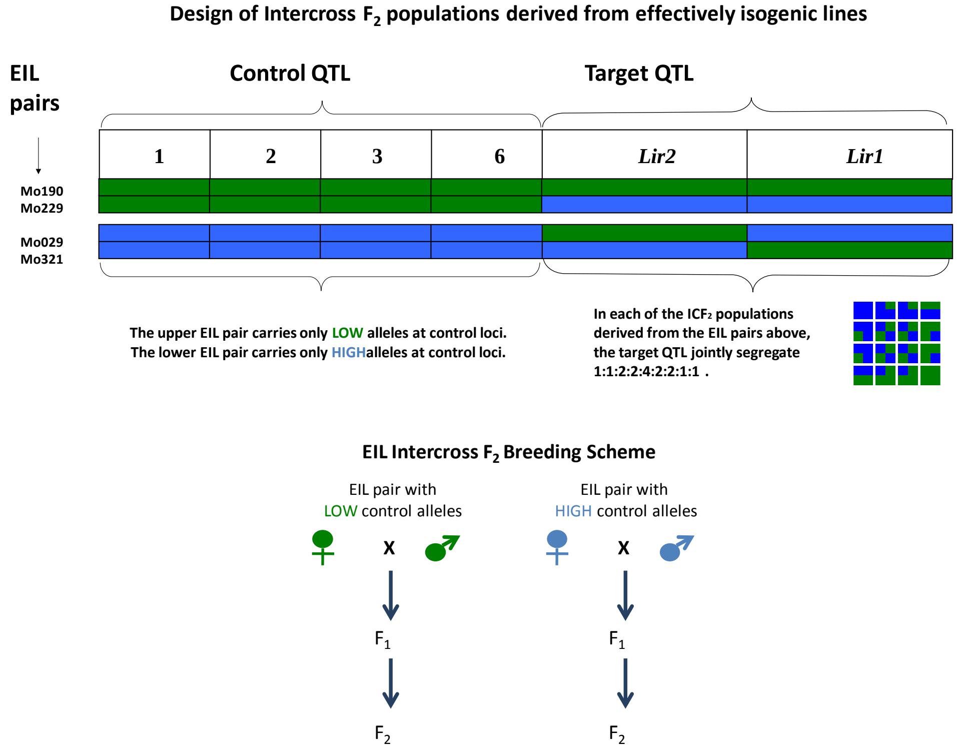
Another caveat is that many independent loci may be specified in a search, which can result in criteria that are so highly defined that no isoline that meets them can be found. In this case, the following error message will be displayed: "No stocks were found with the specified genotype parameters. Please refine your search." In this case however, the search terms will still be summarized on the output page, which will also include the columns of marker names and positions that act as labels for the empty rows of genotype data. As a general rule for genotype data of isolines, one should be able to specify M number of independent (unlinked) genetic markers and produce a successful result so long as $2^{(M)}$ is less than or not grossly greater than the number of isolines being searched. At this time, COGENFITO has no provision to return imperfect matches ranked by how close they come to fully meeting the search criteria, since no scoring system for partial matches has been developed.

# Discussion

The motivation to develop COGENFITO arose from several experimental breeding strategies that have been successfully used (Lauter et al. 2006; Hessel and Lauter, unpublished results). We devote the discussion to describing the use case scenarios for “Intercross $F_2$ analysis of dominance and epitasis” and “Dual testcross QTL analysis”. We also discuss the current limitations of COGENFITO and consider how incremental improvements could make it more useful for commercial breeding applications such as line identification in genomic selection schemes.

*Intercross $F_2$ analysis of dominance and epitasis* Determining the phenotypic consequence of natural allelic variation at particular loci is a common goal of plant breeders and geneticists. Careful use of genetic marker data can reduce the number of generations required to create analysis populations to address this need. When several-to-many QTL affect a trait, accurately assessing the additive, dominance, and interaction effects of a single QTL requires a breeding approach, especially when the QTL effects have initially been characterized using sets of isolines. Traditionally, contrasts of nearly isogenic line (NIL) pairs were used to eliminate variation at non-target loci. We have demonstrated that effectively isogenic line (EIL) pairs can be useful as well (Lauter et al., 2006). EIL pairs fix the alleles at “control loci”, allowing the alleles at “target loci” to vary. Because the use of EILs alleviates the need for introgressions, substantial breeding time is spared (Figure 5).

**Figure 5. Nuanced characterization of dominance and epistatsis**

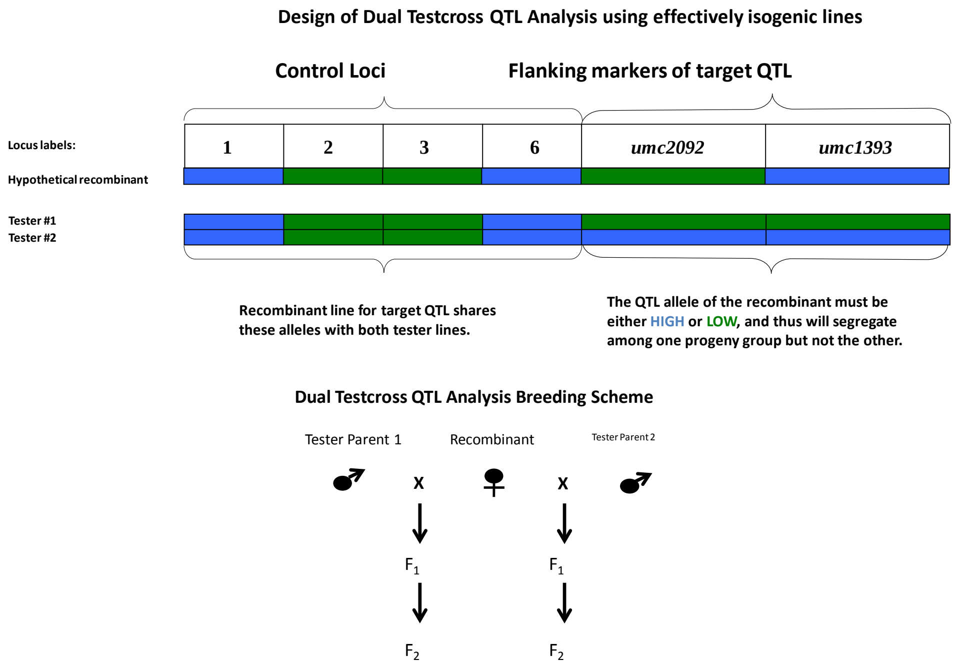

Design of Intercross $F_2$ populations derived from effectively isogenic lines
EIL pairs
Control QTL
Target QTL
1
2
3
6
Lir2
Lir1
Mo190
Mo229
Mo029
Mo321
The upper EIL pair carries only LOW alleles at control loci.
The lower EIL pair carries only HIGHalleles at control loci.
In each of the ICF₂ populations derived from the EIL pairs above, the target QTL jointly segregate 1:1:2:2:4:2:2:1:1 .
EIL Intercross $F_2$ Breeding Scheme
EIL pair with LOW control alleles
EIL pair with HIGH control alleles
x
x
$F_1$
$F_1$
$F_2$
$F_2$

**For each population of $F_2$ plants, phenotype trait of interest and genotype at desired loci. Use ANOVAs to characterize dominance and epistatic interactions.**

COGENFITO is ideal for selecting the EILs to mate in setting up intercross $F_2$ ($ICF_2$) experiments. For the example in Figure 5, several linked markers can be used to specify the QTL genotype at each of the control loci to achieve a “high” background. These ~20 lines (=($[1/2]^4$)*302 RILs) can be further filtered for combinations of contrasting genotypes at the target loci for the potential pairs of EILs. Again, multiple linked markers to ensure correct specification of each QTL genotype should be used for selection. This process would then be repeated for the “low” background. Beyond characterizing the dosage effects of the two QTL and their interactions, interrogation of higher- order epistatic interactions without requiring large population sizes can be achieved with just two generations of breeding (Figure 5).

*Dual testcross QTL analysis* COGENFITO is useful for isoline selection for Dual test-cross QTL analysis (DTQA), an approach that accelerates QTL finemapping for QTLs of moderate effect by permitting allele determination in isolines that harbor informative recombination events. Typically, unless a QTL has a qualitative effect, model-based predictions of whether a particular line carries the high or low QTL allele can’t be trusted, creating a serious limitation for finemapping moderate-effect loci. QTL determination for a recombinant of interest is greatly facilitated by a high signal-to-noise ratio, making moderately penetrant quantitative effects experimentally tractable. The traditional approach to reducing background noise has been to generate NILs, which is laborious as discussed above. Mating pairs of EILs identified using COGENFITO can achieve the same outcome after just two generations of breeding (Figure 6).

**Figure 6.** DTQA anlysis to determine QTL genotype in known recombinants


Design of Dual Testcross QTL Analysis using effectively isogenic lines
Control Loci
Flanking markers of target QTL
Locus labels:
1
2
3
6
umc2092
umc1393
Hypothetical recombinant
Tester #1
Tester #2
Recombinant line for target QTL shares these alleles with both tester lines.
The QTL allele of the recombinant must be either HIGH or LOW, and thus will segregate among one progeny group but not the other.
Dual Testcross QTL Analysis Breeding Scheme
Tester Parent 1
Recombinant
Tester Parent 2
x
x
F1
F1
F2
F2

**Phenotype both progeny groups for trait of interest. Only one of the two $F_2$ groups should segregate for phenotypic variation dependent on the QTL, revealing the QTL genotype of the recombinant line.**

For the DTQA use case, one should first identify the recombinants of interest using COGENFITO as described in the results section. After identifying a recombinant of interest, COGENFITO can be used to browse the genomic regions for each functionally polymorphic locus that must be controlled in order to have effective isogenicity between the tester lines and the isoline harboring the informative recombinantion event (Figure 6). After taking note of the control QTL genotypes needed to match those carried by the line of interest, COGENFITO can be directly used to identify appropriate tester lines with a normal composite genotype search.

*Extending the utility of COGENFITO* The current implementation of COGENFITO at MaizeGDB serves as a useful browser of genotype datasets and greatly facilitates the execution of experimental breeding schemes described in the use cases. However, it has limited utility for application to more complex problems, although these are surmountable by changing existing data types, adding new data types, and using more sophisticated metrics for how matches are determined.

One of the main issues is that the genotype data stored in MaizeGDB for these populations are processed map data, with alternate homozygous classes specified as A and B and missing data denoted by a dash. This nicely accommodates the variety of data types gleaned from SNP, SSLP, CAPS and RFLP marker types, but will not support simple association with allele source after addition of new populations such as the Intermated NC89 x K55 RILs. In this scenario, “A” would mean homozygous B73 in most COGENFITO supported populations, but would mean homozygous NC89 for the INKRILs. Furthermore, markers that distinguish among more than two alleles, such as the 2 to 3% of SNPs that are known to be at least triallelic, are not adequately handled by the present scheme. The main problem is again related to how the data are stored; if the alternative to the B73 allele is not always the same, searches across all populations simultaneously need to be able to distinguish between these second and third alleles. Changing the way the data are stored is readily possible for SNPs, but will require coordination with researchers so that data are delivered in both formats. Since one responsibility of MaizeGDB is to act as a data repository, better support for SNP data should not come at the expense of adequate support for previous data types.

COGENFITO’s functionality could be broadened to allow the inclusion of phenotype data for line selection, and this functionality would likely be of use to plant breeders. Phenotype data have been collected on all 26 populations presently supported, in some cases in multiple environments (Balint-Kurti et al., 2007; Lauter et al., 2008; Buckler et al., 2009; Swanson-Wagner et al., 2009). Handling these data types would require including phenotypic data and descriptions in the selection algorithms, but is a clear possibility. This would permit breeders, for example, to search for lines with particular composite genotypes, while avoiding those with late flowering phenotypes.

Finally, creation of query tools that handle quantitative data would also enable searching according to the extent of the match, rather than according only to whether or not an exact match was found. In crop improvement breeding schemes, it is likely that a breeder would want to optimize combinations of alleles at so many loci that iterative use of COGENFITO in its current form would be required. Since the purposes of COGENFITO are to extend capabilities and reduce the time required to identify lines of interest, the improvements described will be the focus of ongoing development for COGENFITO.

## Methods

*Database Extraction and Technical Details of COGENFITO:* COGENFITO is a built-in database extraction tool that relies on datasets deposited at MaizeGDB. The tool consists of an HTML interface with processing forms that define the genetic mapping population, the genetic markers, and the centiMorgan range. A Javascript-enabled function allows additional loci to be added to the query.

The HTML forms are processed in PHP 4.0 with SQL query statements to access the relevant database tables, which have been previously described (Lawrence et al., 2007). The SQL queries have three primary dependencies. Upon submission of the form, MGDB_MAP_SCORES.NAME is searched for names that contain the markers and mapping population. MAP_SCORES is joined to LOCUS_COORDINATES by PROBED_SITE=ID, and LOCUS_COORDINATES is joined to MAP by MAP=ID. Even though a marker may be found on several different maps, the query is limited by defining which maps in MAP.NAME can be selected. At this point in the database extraction, only one map is found for any given marker. The query then finds the centiMorgan value for the marker on that map and selects the centiMorgan range to return. The final step in the extraction involves returning all MAP_SCORES.SCORES_123 and MAP_SCORES.NAME that fall within that centiMorgan range. A separate built-in function breaks up the SCORES_123 string into positions (individuals) and the positions in the string corresponding to the desired genotype selected are returned along with their formal germplasm names, which are in turn hyperlinked to seed accession pages.

## Acknowledgements

We thank the MaizeGDB team, particularly Darwin Campbell for database support and Carson Andorf for programming support, and the members of the Crop Genomics and Informatics Laboratory database discussion group for their helpful insights. This work was supported by USDA-ARS CRIS Projects 3625-21000-049-00D and 3625-21000-051-00D, as well as by an NSF-IGERT Fellowship awarded to DH through Iowa State University.

## Literature Cited

**Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ** (1990) Basic local alignment search tool. J Mol Biol. **215:** 403-410

**Balint-Kurti PJ, Zwonitzer JC, Wisser RJ, Carson ML, Oropeza-Rosas MA, Holland JB, Szalma SJ** (2007) Precise Mapping of Quantitative Trait Loci for Resistance to Southern Leaf Blight, Caused by Cochliobolus heterostrophus Race O, and Flowering Time Using Advanced Intercross Maize Lines. Genetics **176:** 645-657

**Buckler ES, Holland JB, Bradbury PJ, Acharya CB, Brown PJ, Browne C, Ersoz E, Flint-Garcia S, Garcia A, Glaubitz JC, Goodman MM, Harjes C, Guill K, Kroon DE, Larsson S, Lepak NK, Li H, Mitchell SE, Pressoir G, Peiffer JA, Rosas MO, Rocheford TR, Romay MC, Romero S, Salvo S, Villeda HS, Sofia da Silva H, Sun Q, Tian F, Upadyayula N, Ware D, Yates H, Yu J, Zhang Z, Kresovich S, McMullen MD** (2009) The Genetic Architecture of Maize Flowering Time. Science **325:** 714-718

**Crossett A, Lauter N, Love T** (2010) An Empirical Method for Establishing Positional Confidence Intervals Tailored for Composite Interval Mapping of QTL. PLoS ONE **published 09 Feb 2010**

**Fu Y, Wen T-J, Ronin YI, Chen HD, Guo L, Mester DI, Yang Y, Lee M, Korol AB, Ashlock DA, Schnable PS** (2006) Genetic dissection of intermated recombinant inbred lines using a new genetic map of maize that includes 1,329 novel PCR-based genic markers. Genetics **174:** 1671-1683

**Lauter N, Moscou MJ, Habiger J, Moose SP** (2008) Quantitative Genetic Dissection of Shoot Architecture Traits in Maize: Towards a Functional Genomics Approach. The Plant Genome **1:** 99-110

**Lauter N, Starr M, Moose SP** (2006) High Resolution Mapping and Functional Dissection of QTL Affecting Leaf Epidermal Traits Using the Public IBMRI Maize Population. Proceedings of the Illinois Corn Breeders School **42:** 21-38

**Lawrence CJ, Schaeffer ML, Seigfried TE, Campbell DA, Harper LC** (2007) MaizeGDB's new data types, resources and activities. Nucl. Acids Res. **35:** D895-900

**Lee M, Sharopova N, Beavis WD, Grant D, Katt M, Blair D, Hallauer A** (2002) Expanding the genetic map of maize with the intermated B73 x Mo17 (IBM) population. Plant Mol Biol **48:** 453-461

**Longin C, Utz H, Reif J, Schipprack W, Melchinger A** (2006) Hybrid maize breeding with doubled haploids: I. One-stage versus two-stage selection for testcross performance. TAG Theoretical and Applied Genetics **112:** 903-912

**McMullen MD, Kresovich S, Villeda HS, Bradbury P, Li H, Sun Q, Flint-Garcia S, Thornsberry J, Acharya C, Bottoms C, Brown P, Browne C, Eller M, Guill K, Harjes C, Kroon D, Lepak N, Mitchell SE, Peterson B, Pressoir G, Romero S, Rosas MO, Salvo S, Yates H, Hanson M, Jones E, Smith S, Glaubitz JC, Goodman M, Ware D, Holland JB, Buckler ES** (2009) Genetic Properties of the Maize Nested Association Mapping Population. Science **325:** 737-740

**Prigge V, Melchinger A, Dhillon B, Frisch M** (2009) Efficiency gain of marker-assisted backcrossing by sequentially increasing marker densities over generations. TAG Theoretical and Applied Genetics **119:** 23-32

**Salvi S, Sponza G, Morgante M, Tomes D, Niu X, Fengler KA, Meeley R, Ananiev EV, Svitashev S, Bruggemann E, Li B, Hainey CF, Radovic S, Zaina G, Rafalski JA, Tingey SV, Miao G-H, Phillips RL, Tuberosa R** (2007) Conserved noncoding genomic sequences associated with a flowering-time quantitative trait locus in maize. Proceedings of the National Academy of Sciences **104:** 11376-11381

**Sen TZ, Andorf CM, Schaeffer ML, Harper LC, Sparks ME, Duvick J, Brendel VP, Cannon E, Campbell DA, Lawrence CJ** MaizeGDB becomes 'sequence-centric'. Database **2009:** bap020-

**Sharopova N, McMullen MD, Schultz L, Schroeder S, Sanchez-Villeda H, Gardiner J, Bergstrom D, Houchins K, Melia-Hancock S, Musket T, Duru N, Polacco M, Edwards K, Ruff T, Register JC, Brouwer C, Thompson R, Velasco R, Chin E, Lee M, Woodman-Clikeman W, Long MJ, Liscum E, Cone K, Davis G, Coe EH, Jr.** (2002) Development and mapping of SSR markers for maize. Plant Mol Biol **48:** 463-481

**Swanson-Wagner RA, DeCook R, Jia Y, Bancroft T, Ji T, Zhao X, Nettleton D, Schnable PS** (2009) Paternal Dominance of Trans-eQTL Influences Gene Expression Patterns in Maize Hybrids. Science **326:** 1118-1120