MODEM maize Multidimensional Omics Data Envelopment & Mining database

Manual for MODEM database

Updated on 11/10/2015, Sunday

This manual is created with the expectation to help you better understand the database of MODEM, including simple introduction of the background of MODEM created for, how MODEM can indeed help in your daily researches, several examples to illustrate how to realize them, and at the end of this document, the full description on materials collection and data generation/processing/evaluation are provided.

MODEM has been published on *XXX*, and the latest version was released at 11/10/2015. If you've used MODEM in your work, please cite in your own publications as:

Any comments or problems please contact:

Dr. Jianxiao Liu (liujianxiao@mail.hzau.edu.cn) or
Mr. Haijun Liu (heroalone@webmail.hzau.edu.cn).

What are the main goals for MODEM?

Recent years, the related maize multidimensional omics data got rapid growth. These big data made bright promising, especially provide great advantages via association mapping, to better understand the genetic artitecture of specific phenotype, and in return to deep annotate the complex genome. While the researches of genetics and genomics are progressing rapidly in maize, however, the resources of related database fall behind and are limited for next analyses.

The MODEM database is thus developed to meet most of these goals. With the help of widely collaboration, MODEM has integrated abundant datasets derived from the same diverse panel, including variations from genome, transcriptome, metabolome and phenome, and related association mapping results. MODEM would be mainly focused on functional genome of maize, especially on exploring the regulatory mechanism of related (putative) causal QTL/gene.

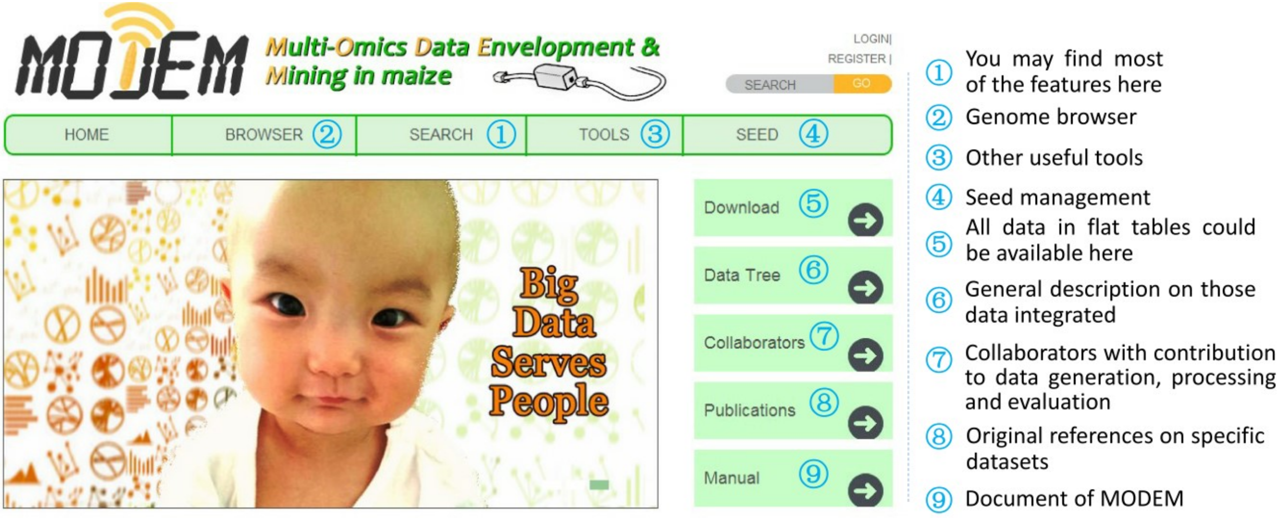
What are the types of data integrated in MODEM?

As we have collected 527 maize elite inbred lines and generated multiple omics-level maize bio-data and the consequent mapping results. The general description of multi-omics data involved in MODEM are listed as below, more details (Overall Data Structure) could be available at the data-tree page: http://modem.hzau.edu.cn/maizego/edit.jsp

Index	Data
Germplasm Resources	527 inbreds for association mapping panel (AMP) with different populations (143 lines for NSS, non-stiff-stock; 33 for SS, Stiff-stock; 232 for TST, Tropical and Semi-tropical; and the left 119 are regared as MIXED)
Genomic variation	~50K SNPs from MaizeSNP50 BeadChip for AMP (513 lines), of which 368 have >1.03 million SNPs by RNA-Seq with 0.56 million passed the MAF > 0.05 filtering. The whole 513 panel is finally imputated to 0.56M SNPs
Transcriptome quantification	28,769 genes' quantitative expression of maize whole kernel (15 days after pollination, 15 DAP)
Phenotype measurement	nearly agronomic 50 traits includingyield, response to drought, floods and diseases with 4-8 locations and multiple years (ranging from 2007 to 2012) of the whole AMP
metabolomics	983 metabolic profiling of AMP and 17 amino acid components identified within 2 location for AMP
mapping results	Mapping results of association analysis for different traits (including expression levels)

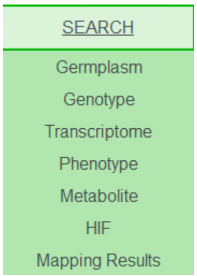
Features list of MODEM and examples to realize it step-by-step

A general view on functions of MODEM

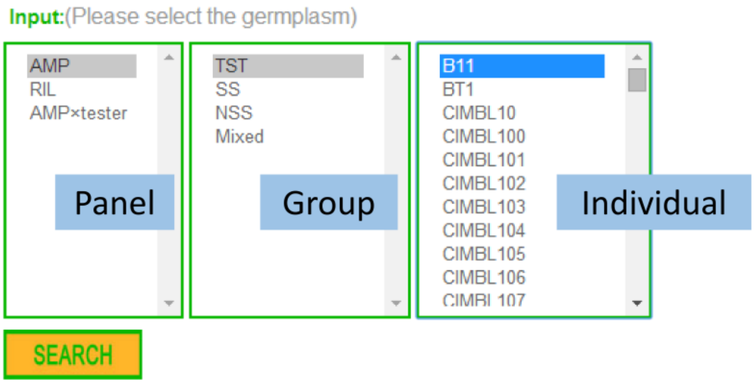


① You may find most of the features here
② Genome browser
③ Other useful tools
④ Seed management
⑤ All data in flat tables could be available here
⑥ General description on those data integrated
⑦ Collaborators with contribution to data generation, processing and evaluation
⑧ Original references on specific datasets
⑨ Document of MODEM

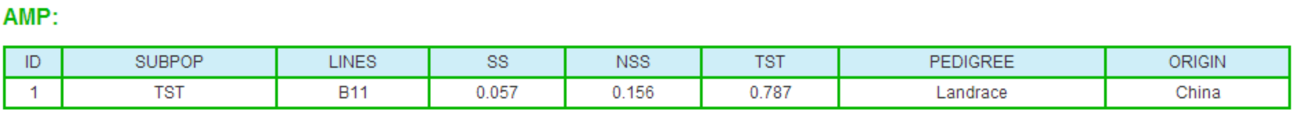
The search menu (①)



"Germplasm": This is used to display specific information of germplasms. From here, you could easily get a general view of those materials used for the most of implements built in MODEM.



You can select the Panel and Group to see individuals under them, and select each individual to learn more details about the sub-population (group), structure component, pedigree and its origination:

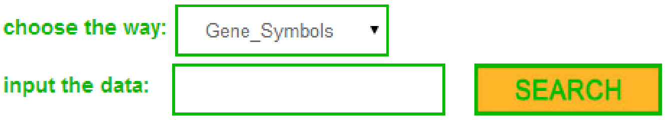


AMP:

ID	SUBPOP	LINES	SS	NSS	TST	PEDIGREE	ORIGIN
1	TST	B11	0.057	0.156	0.787	Landrace	China

If no individual selected, the lines under the whole group would be displayed.

“Genotype”: This is used to search for genotypes (filtered with MAF<5%) derived from interested region.

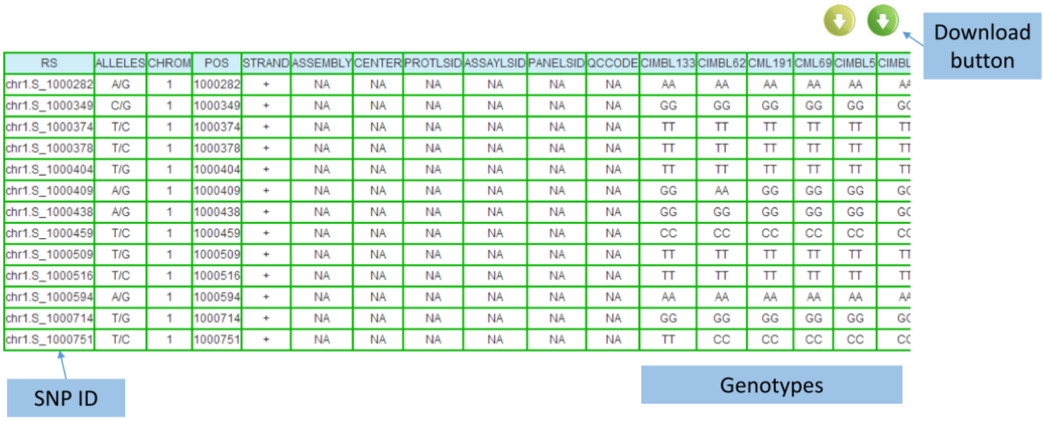


choose the way: Gene_Symbols

input the data:

Two way are provided to search: through gene symbol or specific physical region. Gene symbols are gene IDs named as Ensemble regulation, such GRMZM2G374777, and the physical region should be typed as chromosome:start-position..end-position, just like chr1:100339360..10037221.

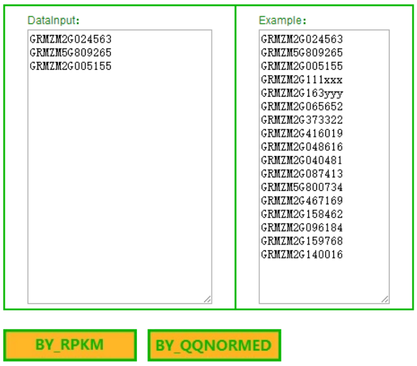
The SNPs located within the selected gene or related region would be displayed in HAPMAP format. For SNPs from RNA-seq, the ID is uniquely labeled as chr1.S_1000282, while “chr” represents the “chromosome”, the “S” means “SNP” and the “1000282” is physical coordinate of this variation, and for SNPs from MaizeSNP50 BeadChip, the ID remains as from the Chip. This list could also be directly saved to local files with TXT or CSV formats for further analysis.



RS	ALLELES	CHROM	POS	STRAND	ASSEMBLY	CENTER	PROTLSID	ASSAYLSID	PANELSID	QCCODE	CIMBL133	CIMBL62	CML191	CML69	CIMBL5	CIMBL
chr1.S_1000282	A/G	1	1000282	+	NA	NA	NA	NA	NA	NA	AA	AA	AA	AA	AA	AA
chr1.S_1000349	C/G	1	1000349	+	NA	NA	NA	NA	NA	NA	GG	GG	GG	GG	GG	GG
chr1.S_1000374	T/C	1	1000374	+	NA	NA	NA	NA	NA	NA	TT	TT	TT	TT	TT	TT
chr1.S_1000378	T/C	1	1000378	+	NA	NA	NA	NA	NA	NA	TT	TT	TT	TT	TT	TT
chr1.S_1000404	T/G	1	1000404	+	NA	NA	NA	NA	NA	NA	TT	TT	TT	TT	TT	TT
chr1.S_1000409	A/G	1	1000409	+	NA	NA	NA	NA	NA	NA	GG	AA	GG	GG	GG	GG
chr1.S_1000438	A/G	1	1000438	+	NA	NA	NA	NA	NA	NA	GG	GG	GG	GG	GG	GG
chr1.S_1000459	T/C	1	1000459	+	NA	NA	NA	NA	NA	NA	CC	CC	CC	CC	CC	CC
chr1.S_1000509	T/G	1	1000509	+	NA	NA	NA	NA	NA	NA	TT	TT	TT	TT	TT	TT
chr1.S_1000516	T/C	1	1000516	+	NA	NA	NA	NA	NA	NA	TT	TT	TT	TT	TT	TT
chr1.S_1000594	A/G	1	1000594	+	NA	NA	NA	NA	NA	NA	AA	AA	AA	AA	AA	AA
chr1.S_1000714	T/G	1	1000714	+	NA	NA	NA	NA	NA	NA	GG	GG	GG	GG	GG	GG
chr1.S_1000751	T/C	1	1000751	+	NA	NA	NA	NA	NA	NA	TT	CC	CC	CC	CC	CC

“Transcriptome”: You could get genes’ expression quantification (15DAP) for the whole AMP panel here. However, only genes with filtered set and expressed at larger than 50% lines (28,769 genes left) were retained. The expression values by RPKM normalization or further normal quantile transformation (Q-Q normed; more details described at the end of this manual) could be displayed and saved when a list of gene symbols provided.

The whole list (38,850 genes, expressed in at least one genotype) could be downloaded on the DOWNLOAD page.



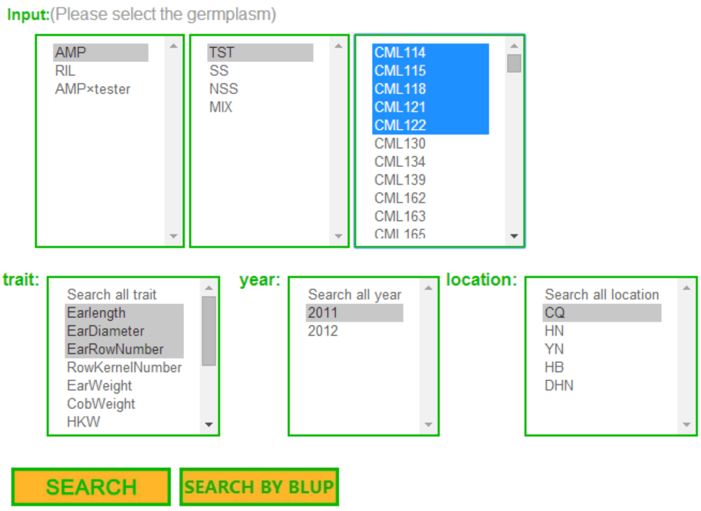
DataInput:

```
GRMZM2G024563
GRMZM5G809265
GRMZM2G005155
```

Example:

```
GRMZM2G024563
GRMZM5G809265
GRMZM2G005155
GRMZM2G111xxx
GRMZM2G163yyy
GRMZM2G065652
GRMZM2G373322
GRMZM2G416019
GRMZM2G048616
GRMZM2G040481
GRMZM2G087413
GRMZM5G800734
GRMZM2G467169
GRMZM2G158462
GRMZM2G096184
GRMZM2G159768
GRMZM2G140016
```

“Phenotype”: In this interface, users can fetch phenotypes of different populations or accessions through selecting trait name, interested germplasm, and corresponding year and location. People could hold down the CTRL key to multi-select each option, and choose to display specific original phenotypes or to make BLUP calculation.



Input:(Please select the germplasm)

- AMP
- RIL
- AMP×tester

- TST
- SS
- NSS
- MIX

- CML114
- CML115
- CML118
- CML121
- CML122
- CML130
- CML134
- CML139
- CML162
- CML163
- CML165

trait:

- Search all trait
- Earlength
- EarDiameter
- EarRowNumber
- RowKernelNumber
- EarWeight
- CobWeight
- HKW

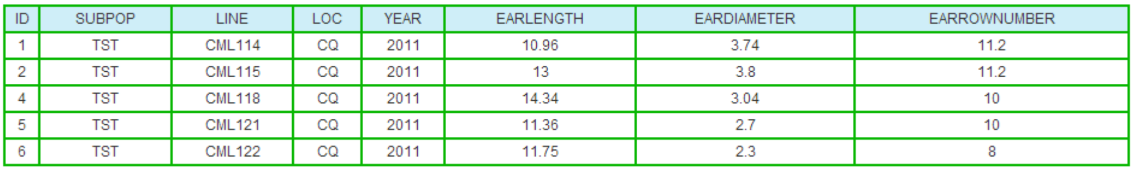
year:

- Search all year
- 2011
- 2012

location:

- Search all location
- CQ
- HN
- YN
- HB
- DHN

The above indicate to search three ear traits (including Ear length, Ear Diameter, number of Ear Row) at year 2011 and Chongqing (CQ) location from five individuals (belong to AMP and TST group), and corresponding search results are:

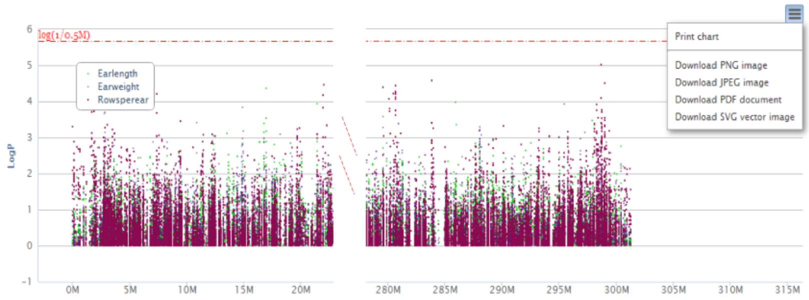


ID	SUBPOP	LINE	LOC	YEAR	EARLENGTH	EARDIAMETER	EARROWNUMBER
1	TST	CML114	CQ	2011	10.96	3.74	11.2
2	TST	CML115	CQ	2011	13	3.8	11.2
4	TST	CML118	CQ	2011	14.34	3.04	10
5	TST	CML121	CQ	2011	11.36	2.7	10
6	TST	CML122	CQ	2011	11.75	2.3	8

【**NOTICE:** The number of each phenotype were real appraisal values (average form 3-5 individuals/repeates), which may not meet normal distribution.】

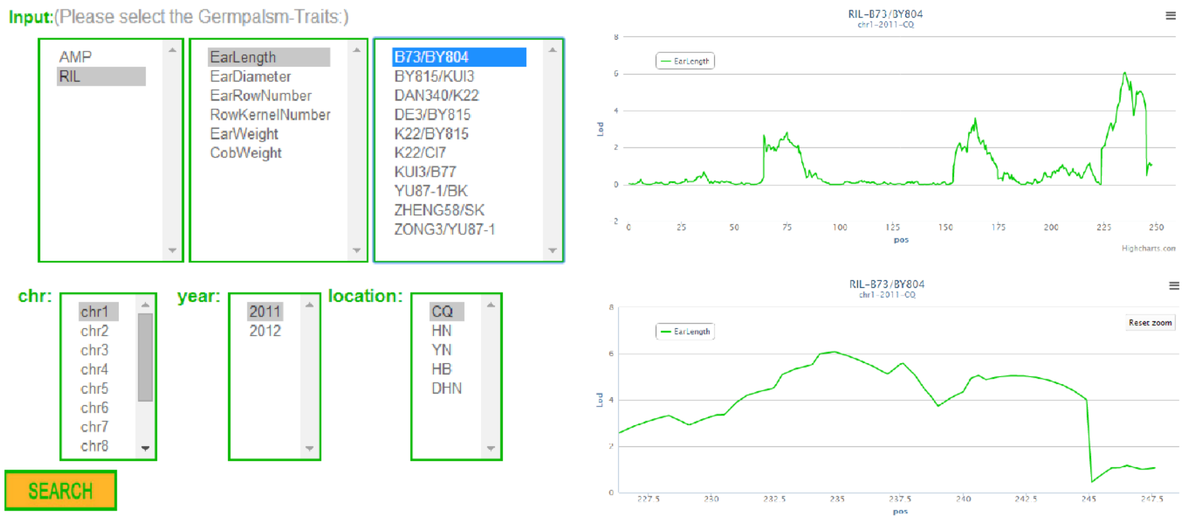
Researchers could carry out mapping analysis themselves after downloading the phenotypes (including BLUPed ones), or search and download the pre-existing mapping results directly from the database.

"Mapping results": Results of both association and linkage mapping for selected traits are provided. Right now only the ear related traits are available, and association mapping results are from the BLUPed traits while the linkage mapping results could be displayed by separated year and locations.



Traits could be multi-selected by holding CTRL key.

Left shows the association mapping results on three traits including Ear length, Ear weight and Row number per ear. The plot could be saved in multiple image formats.

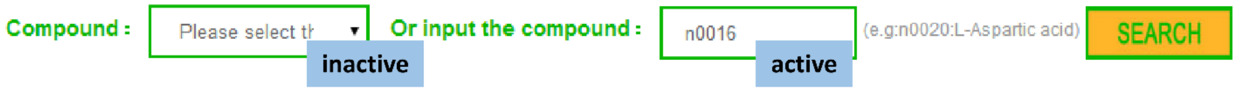


The above displays the linkage mapping results of Ear length trait on B73/BY804 RIL population at year 2011 and CQ location. As the interactive JavaScript technology (HIGHCHART, http://www.highcharts.com/) is applied, users could zoom in interested regions on the plot to see much more details or saved to image files. The picture at bottom right corner of above shows the third peak of the results.

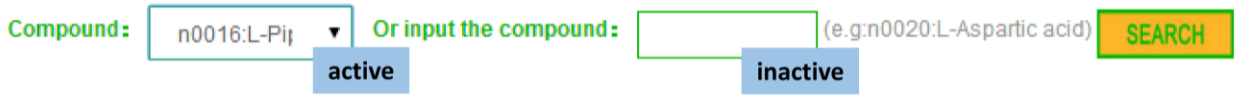
“Metabolite”: Based on LC-MS/MS analysis we have quantified 983 metabolite features on the AMP lines (with three environments, named as E1, E2, E3) and two RIL (B73 × BY804, BB and ZONG3 × YU87-1, ZY) populations. From the “List of 983 metabolite features” button, you could learn (or download) the list of metabolites and detailed information.

As metabolite-based genome-wide association studies (mGWAS) and linkage mapping are conducted in previous study, genome-wide significant loci are also identified. Detailed mapping results and putative candidates can be accessed (or downloaded) through “Significant loci identified by GWAS”, “QTL mapping summary” and “List of possible candidate genes”, respectively.

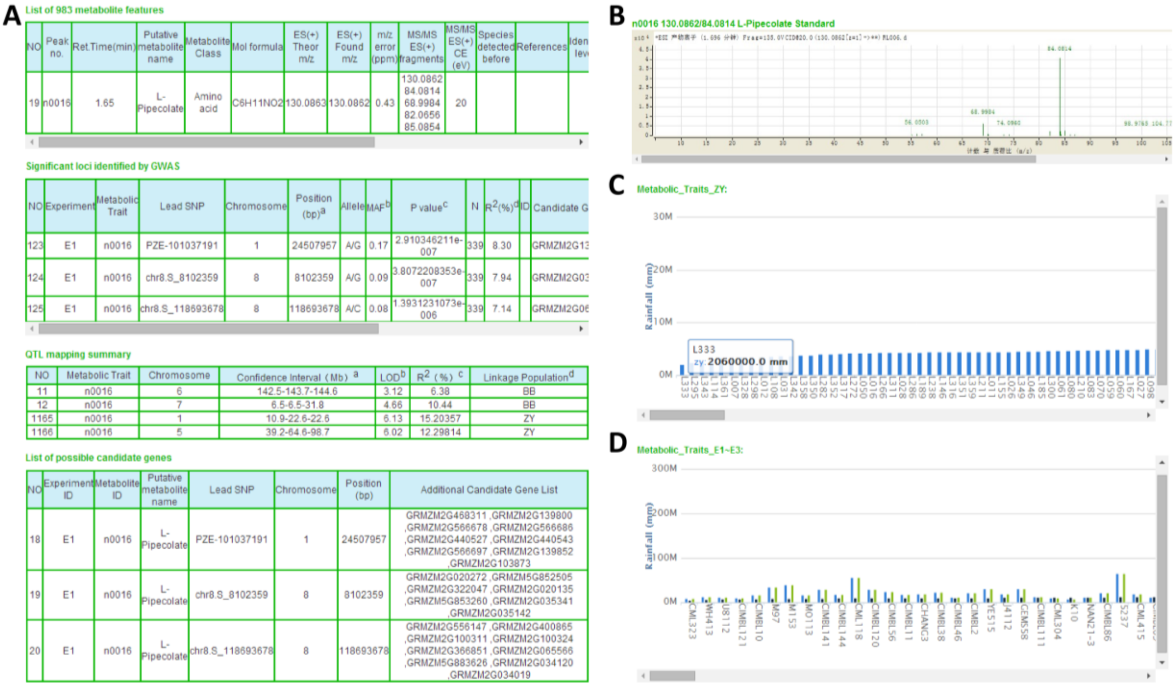
However, we have integrated all above information together through selecting (or inputting) the interested metabolite ID. 【NOTICE: the selecting and inputting could not simultaneous action.】


Compound : Please select th ▾ inactive
Or input the compound : n0016 active

Or


Compound: n0016:L-Pi ▾ active
Or input the compound: inactive

Taking the n0016 (L-Pipecolate, belongs to amino acid) as example, either of the two above is worked, and the searching integrated results are displayed as:



List of 983 metabolite features

NO	Peak no.	Ret.Time(min)	Putative metabolite name	Metabolite Class	Mol formula	ES(+) Theor m/z	ES(+) Found m/z	m/z error (ppm)	MS/MS ES(+) fragments	MS/MS ES(+) CE (eV)	Species detected before	References	Iden lev
19	n0016	1.65	L-Pipecolate	Amino acid	C6H11NO2	130.0863	130.0862	0.43	130.0862 84.0814 68.9984 82.0656 85.0854	20			

Significant loci identified by GWAS

NO	Experiment	Metabolic Trait	Lead SNP	Chromosome	Position (bp)^a	Allele	MAF^b	P value^c	N	R^2(%)^d	ID	Candidate G
123	E1	n0016	PZE-101037191	1	24507957	A/G	0.17	2.910346211e-007	339	8.30		GRMZM2G13
124	E1	n0016	chr8.S_8102359	8	8102359	A/G	0.09	3.8072208353e-007	339	7.94		GRMZM2G03
125	E1	n0016	chr8.S_118693678	8	118693678	A/C	0.08	1.3931231073e-006	339	7.14		GRMZM2G06

QTL mapping summary

NO	Metabolic Trait	Chromosome	Confidence Interval（Mb）^{a}	LOD^b	R^2（%）^{c}	Linkage Population^d
11	n0016	6	142.5-143.7-144.6	3.12	6.38	BB
12	n0016	7	6.5-6.5-31.8	4.66	10.44	BB
1165	n0016	1	10.9-22.6-22.6	6.13	15.20357	ZY
1166	n0016	5	39.2-64.6-98.7	6.02	12.29814	ZY

List of possible candidate genes

NO	Experiment ID	Metabolite ID	Putative metabolite name	Lead SNP	Chromosome	Position (bp)	Additional Candidate Gene List
18	E1	n0016	L-Pipecolate	PZE-101037191	1	24507957	GRMZM2G468311 ,GRMZM2G139800 ,GRMZM2G566678 ,GRMZM2G566686 ,GRMZM2G440527 ,GRMZM2G440543 ,GRMZM2G566697 ,GRMZM2G139852 ,GRMZM2G103873
19	E1	n0016	L-Pipecolate	chr8.S_8102359	8	8102359	GRMZM2G020272 ,GRMZM5G852505 ,GRMZM2G322047 ,GRMZM2G020135 ,GRMZM5G853260 ,GRMZM2G035341 ,GRMZM2G035142
20	E1	n0016	L-Pipecolate	chr8.S_118693678	8	118693678	GRMZM2G556147 ,GRMZM2G400865 ,GRMZM2G100311 ,GRMZM2G100324 ,GRMZM2G366851 ,GRMZM2G065566 ,GRMZM5G883626 ,GRMZM2G034120 ,GRMZM2G034019

n0016 130.0862/84.0814 L-Pipecolate Standard

x10⁴
+ESI 产物离子 (1.696 分钟) Frag=135.0V CID@20.0 (130.0862[z=1] -> **) RL006.d
84.0814
68.9984
56.0503
74.0960
98.9765
104.77
计数 与 质荷比 (m/z)

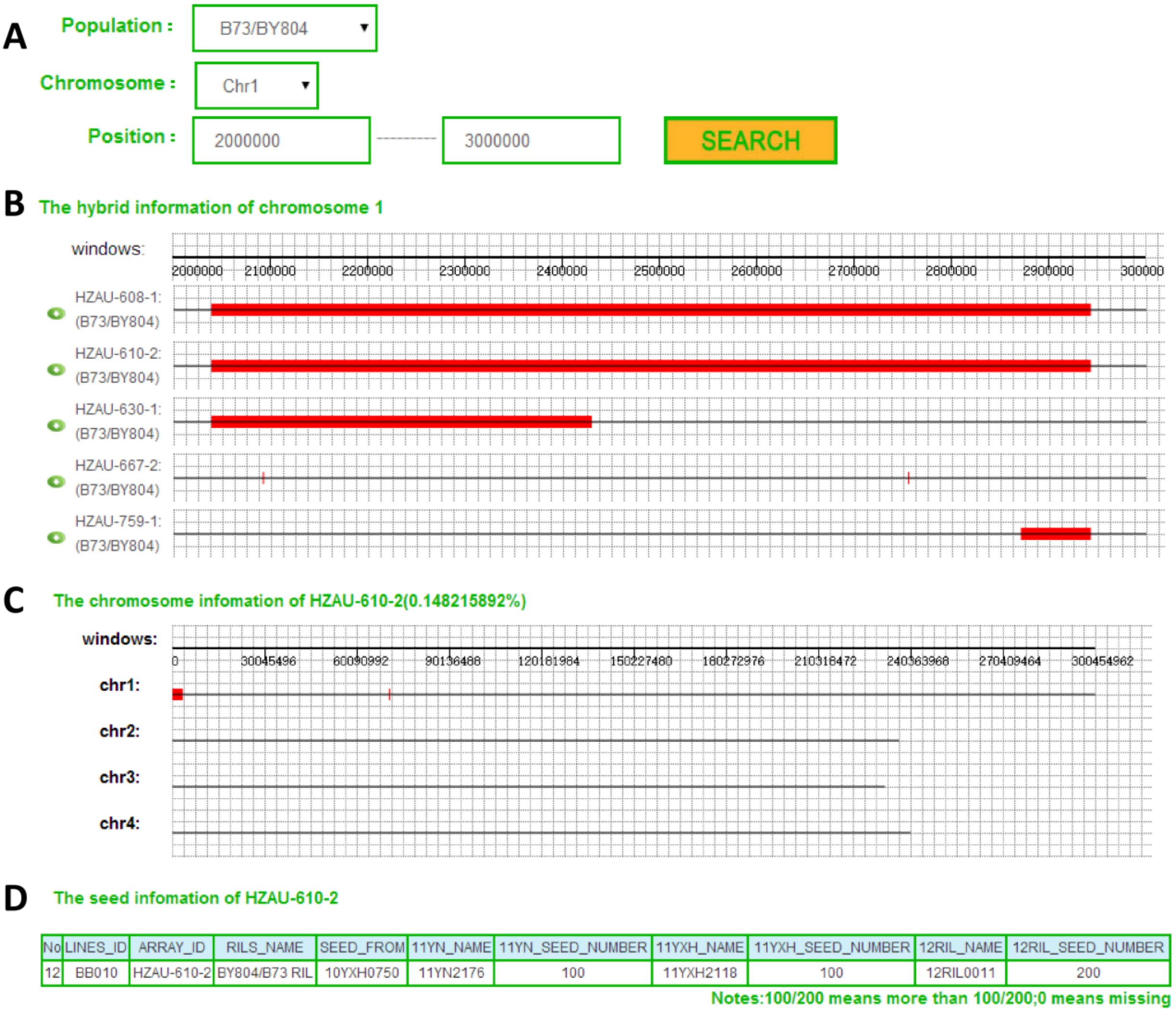
Metabolic_Traits_ZY:

Metabolic_Traits_E1~E3:

(A) List of the detailed features, results derived from association and linkage mapping, and the putative candidates. (B) displays the pre-stored mass spectrum graph of selected compound. (C) and (D) indicate the phenotypic distribution among RILs and AMP populations, while the lines are ordered as their observed trait values (either one of the multiple environments measured for AMP). The exact phenotypic value could be available for each individual when moved mouse over.

"HIF": Heterogeneous inbred family (HIF), represents those individuals who own heterogeneous segment(s) even after self-crossed for several generations, and could be successfully applied to QTL fine mapping. With the help of developed 12 RILs by single seed descent and the construction of ultrahigh-density linkage maps, we identified more than 16,662 heterogeneous segments (Liu et al., prepared) under an acceptable screening process, with an average of 1,388 per population (ranging from 395 to 1,783) and 7.6 per line. Of which, up to 33.8% is shorter than 2M, more than 90% exists in all populations, and, except two RILs, more than four lines on average shared each heterogeneous segment, suggesting the high resolution and coverage of the heterogeneous segments, which make it an ideal resource to QTL fine mapping.

Each individual that own heterogeneous segment(s) (HS, colored in red) within the searching interval in selected population would be displayed. Below is an example for searching HIF from BB RIL population during region from 2M to 3M on chromosome1 (A), and those individuals with HS during target region are displayed (B, only subset are showed here). Clicking the download icon on the left of each line would obtain the individual's whole genotype information, with number 1 and 2 represent two alleles from different parents and number 3 means heterogeneours. While clicking on the line name would result to display its HS information on whole genome level, as (C) indicates the genome-wide HS distribution (partially displayed here) and the heterogeneous rate (0.148%) of line HZAU-610-2. Researchers could further see the seed storage information and require the seed with heterogeneours segments in interested region and promote their specific studies (D).



A

Population : B73/BY804

Chromosome : Chr1

Position : 2000000 --------- 3000000

B The hybrid information of chromosome 1

C The chromosome infomation of HZAU-610-2(0.148215892%)

D The seed infomation of HZAU-610-2

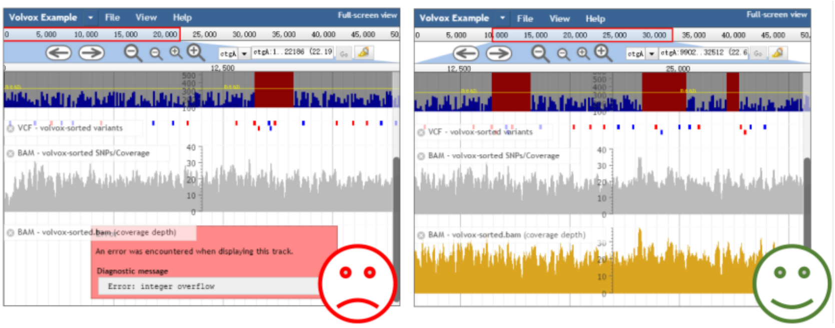
No	LINES_ID	ARRAY_ID	RILS_NAME	SEED_FROM	11YN_NAME	11YN_SEED_NUMBER	11YXH_NAME	11YXH_SEED_NUMBER	12RIL_NAME	12RIL_SEED_NUMBER
12	BB010	HZAU-610-2	BY804/B73 RIL	10YXH0750	11YN2176	100	11YXH2118	100	12RIL0011	200

Notes:100/200 means more than 100/200;0 means missing

Genome *BROWSER* (②) for displaying genotype and expression variation revealed by RNA-seq. The original sequencing reads and variation both in qualitative and quantitative were both involved. Kinds of tracks could be selected to display and are listed in following table:

Category Name	Track included	Description
Reference sequence	Reference sequence	Reference sequence with amino acids from six possible reading frames
Miscellaneous	GFF3	Gene structure annotation, expression comparison or tissue-based expression patterns
SNP	SNPs from MaizeGo lab	SNP information including ID, effect type, eQTL, allele frequency and alleles for each individual
Original reads	B73-sorted BAM	Sequences from original reads covered within selected regions
Coverage	B73-sorted Coverage	Reads coverage for specific region in histogram
Quantitative Density/XYPlot	BigWig Density/XYPlot	Reads coverage for specific region in density plot or XYPlot

【**NOTICE**: Jbrowse is used to build the genome browser, and some web-browsers may fail to display several tracks correctly, for example the "integer overflow" error. The official reminding is that "*BAM, VCF, and BigWig tracks not available on Internet Explorer 9 and earlier*". However, "**Google Chrome**", **"Mozilla Firefox", "Apple Safari" and "Opera"** are all successfully tested (on Window 8.1 64-bit). And we highly suggest users to firstly load http://jbrowse.org/ to see if it works well as showed below for the two different situations.】

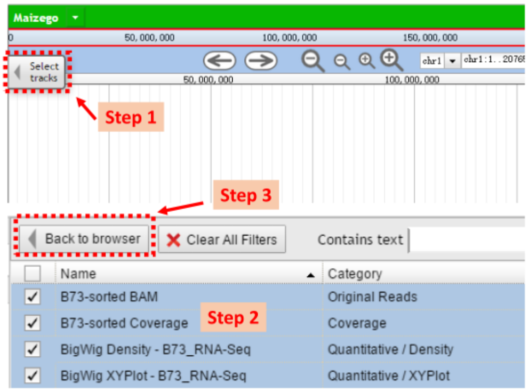

Volvox Example
File
View
Help
Full-screen view
VCF - volvox-sorted variants
BAM - volvox-sorted SNPs/Coverage
BAM - volvox-sorted.bam (coverage depth)

An error was encountered when displaying this track.

Diagnostic message

Error: integer overflow

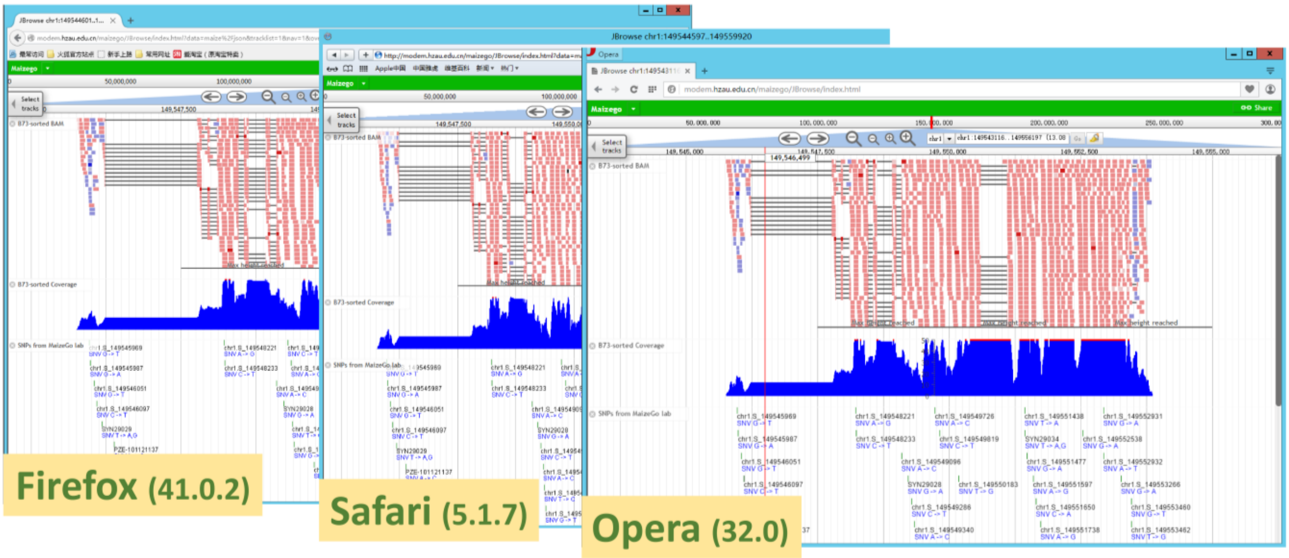
The first time to load the page may show like left below (up-panel), which represents there's no tracks selected to display. Click "**Select tracks**" (Step 1), the tracks listed above would appear to be selected. After selecting interested tracks (Step 2), click "Back to browser" button (Step 3) and those information belong to specific track would be displayed. And the order (from up to bottom) of tracks is the same with the selected order in Step 2, however, users could easily change the order by moving the track name.



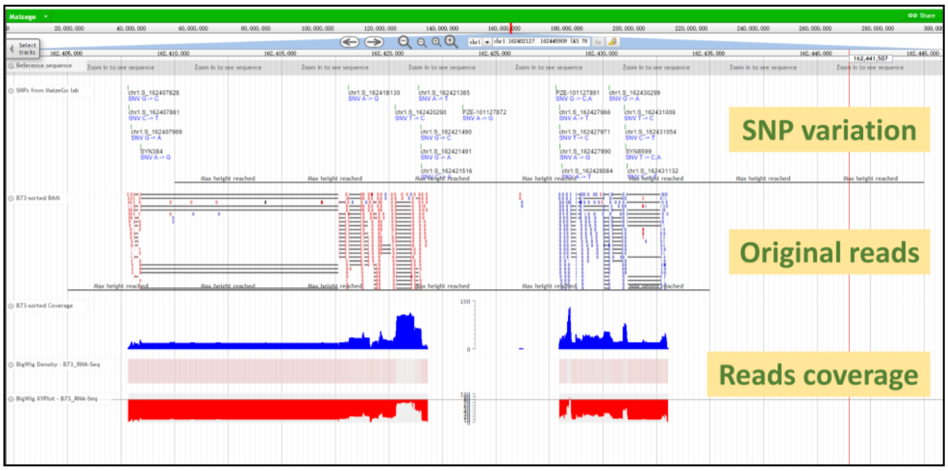
☐	Name	Category
☑	B73-sorted BAM	Original Reads
☑	B73-sorted Coverage	Coverage
☑	BigWig Density - B73_RNA-Seq	Quantitative / Density
☑	BigWig XYPlot - B73_RNA-Seq	Quantitative / XYPlot

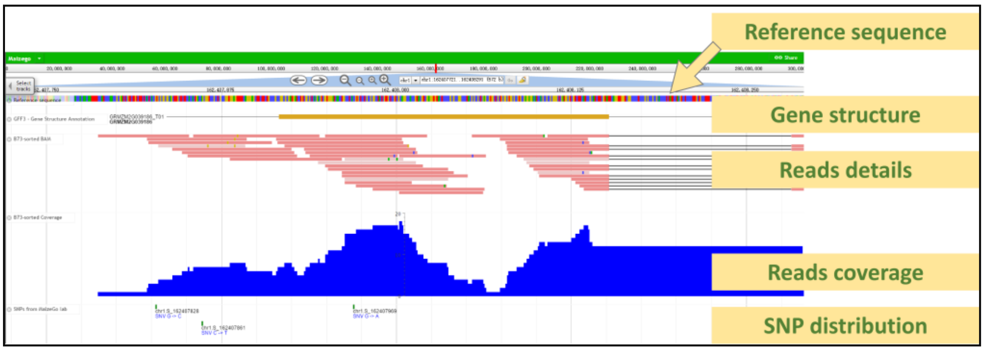
【**NOTICE**: Because right now we have too much data integrated in the Browser, it may need large memory (>2G) that would be out of run in PC. Thus we highly suggest to **only select interested tracks** to display, **or several tracks with limited region (<100Kb)**】

Here are the examples for different web-browsers to correctly display the genome browser, and following cases are all displayed by Google Chrome.



The common used displaying tracks are like these:

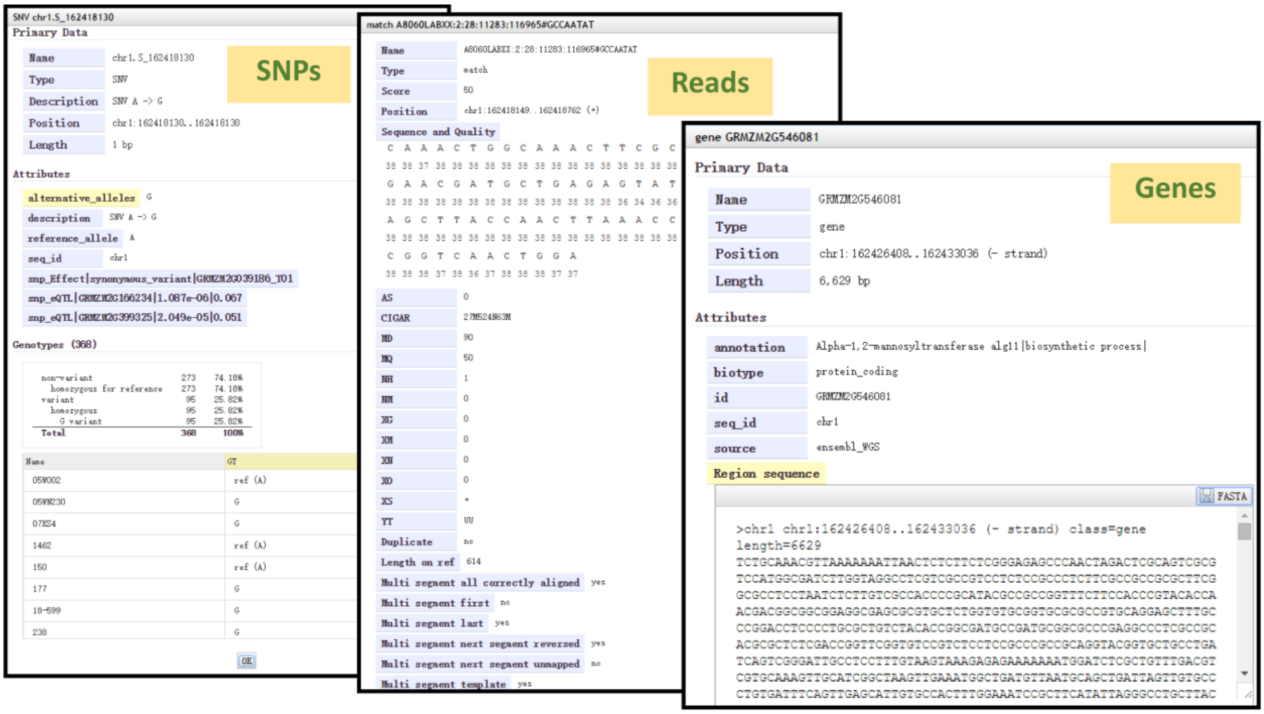




Details displaying:

By left-clicking on specific tracks, people could get much more details. For examples below, we can get information for SNP variation including SNP ID, SNP alleles, SNP effect (formatted as “Type|target gene”, such as synonymous variant|GRMZM2G039186_T01 for SNP chr1.S_162418130, which means the SNP is a synonymous variant for the 1st isoform of gene GRMZM2G039186), eQTL mapping information (formatted as “Targeted gene|P-value|R-square” which calculated by mixed liner model on single marker), SNP frequency and exact allele for every individual; for original reads including reads ID, sequence and its quality for each base and mapping statistics; for genes including Gene symbol, physical position, length, fully annotation (combined from almost all available resources including classical names, predictive function, GO description, encoded enzyme, involved metabolic pathways et al.) and sequences for whole gene or CDS et al.

People may have their interested candidates, and the SNP variation and its eQTL target could provide potential molecular mechanisms and the original reads for different individuals could be applied in designing primers.



SNV chr1.S_162418130

Primary Data

Name	chr1.S_162418130
Type	SNV
Description	SNV A -> G
Position	chr1:162418130..162418130
Length	1 bp

Attributes

alternative_alleles	G
description	SNV A -> G
reference_allele	A
seq_id	chr1
snp_Effect|synonymous_variant|GRMZM2G039186_T01	
snp_eQTL|GRMZM2G166234|1.087e-06|0.067	
snp_eQTL|GRMZM2G399325|2.049e-05|0.051	

Genotypes (368)

```
non-variant                  273   74.18%
  homozygous for reference   273   74.18%
variant                       95   25.82%
  homozygous                  95   25.82%
    G variant                 95   25.82%
Total                        368    100%
```

Name	GT
05W002	ref (A)
05WN230	G
07KS4	G
1462	ref (A)
150	ref (A)
177	G
18-599	G
238	G

match A8060LABXX:2:28:11283:116965#GCCAATAT

Name	A8060LABXX:2:28:11283:116965#GCCAATAT
Type	match
Score	50
Position	chr1:162418149..162418762 (+)

Sequence and Quality

```
C  A  A  A  C  T  G  G  C  A  A  A  C  T  T  C  G  C
38 38 37 38 38 38 38 38 38 38 38 38 38 38 38 38 38 38
G  A  A  C  G  A  T  G  C  T  G  A  G  A  G  T  A  T
38 38 38 38 38 38 38 38 38 38 38 38 38 38 36 34 36 36
A  G  C  T  T  A  C  C  A  A  C  T  T  A  A  A  C  C
38 38 38 38 38 38 38 38 38 38 38 38 38 38 38 38 38 38
C  G  G  T  C  A  A  C  T  G  G  A
38 38 38 37 38 36 37 38 38 38 37 37
```

AS	0
CIGAR	27M524N63M
MD	90
MQ	50
NH	1
NM	0
XG	0
XM	0
XN	0
XO	0
XS	+
YT	UU
Duplicate	no
Length on ref	614
Multi segment all correctly aligned	yes
Multi segment first	no
Multi segment last	yes
Multi segment next segment reversed	yes
Multi segment next segment unmapped	no
Multi segment template	yes

gene GRMZM2G546081

Primary Data

Name	GRMZM2G546081
Type	gene
Position	chr1:162426408..162433036 (- strand)
Length	6,629 bp

Attributes

annotation	Alpha-1,2-mannosyltransferase alg11|biosynthetic process|
biotype	protein_coding
id	GRMZM2G546081
seq_id	chr1
source	ensembl_WGS

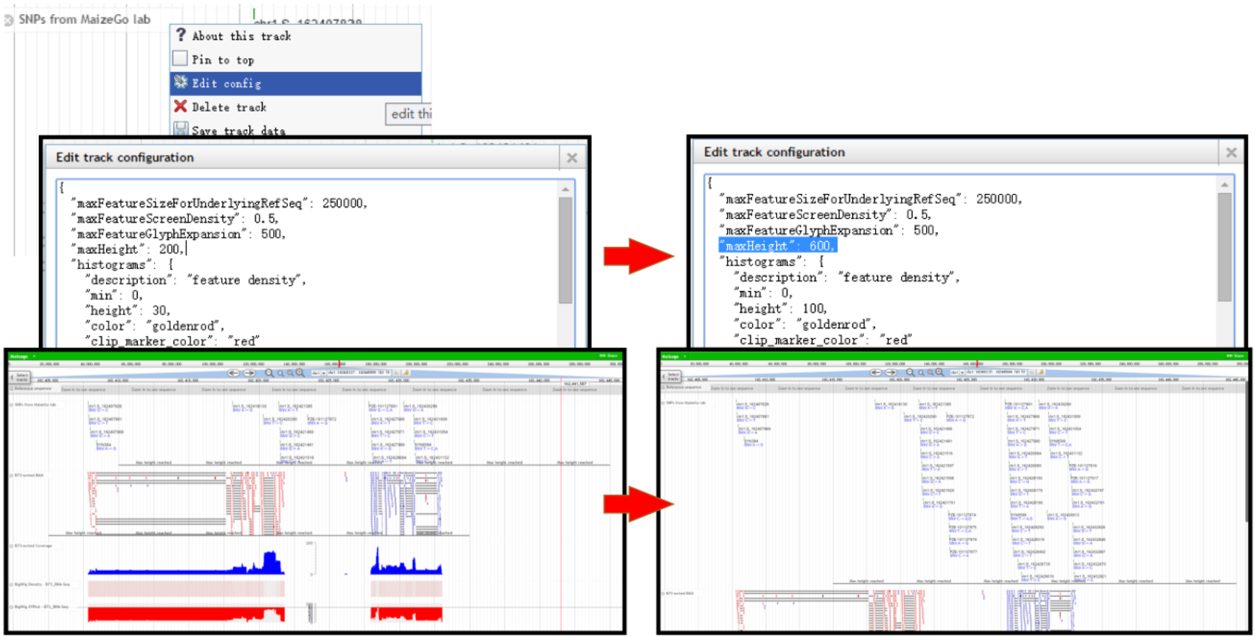
Region sequence

FASTA

```
>chr1 chr1:162426408..162433036 (- strand) class=gene
length=6629
TCTGCAAACGTTAAAAAAATTAACTCTCTTCTCGGGAGAGCCCAACTAGACTCGCAGTCGCG
TCCATGGCGATCTTGGTAGGCCTCGTCGCCGTCCTCTCCGCCCTCTTCGCCGCCGCGCTTCG
GCGCCTCCTAATCTCTTGTCGCCACCCCGCATACGCCGCCGGTTTCTTCCACCCGTACACCA
ACGACGGCGGCGGAGGCGAGCGCGTGCTCTGGTGTGCGGTGCGCGCCGTGCAGGAGCTTTGC
CCGGACCTCCCCTGCGCTGTCTACACCGGCGATGCCGATGCGGCGCCCGAGGCCCTCGCCGC
ACGCGCTCTCGACCGGTTCGGTGTCCGTCTCCTCCGCCCGCCGCAGGTACGGTGCTGCCTGA
TCAGTCGGGATTGCCTCCTTTGTAAGTAAAGAGAGAAAAAAAATGGATCTCGCTGTTTGACGT
CGTGCAAAGTTGCATCGGCTAAGTTGAAATGGCTGATGTTAATGCAGCTGATTAGTTGTGCC
CTGTGATTTCAGTTGAGCATTGTGCCACTTTGGAAATCCGCTTCATATTAGGGCCTGCTTAC
```

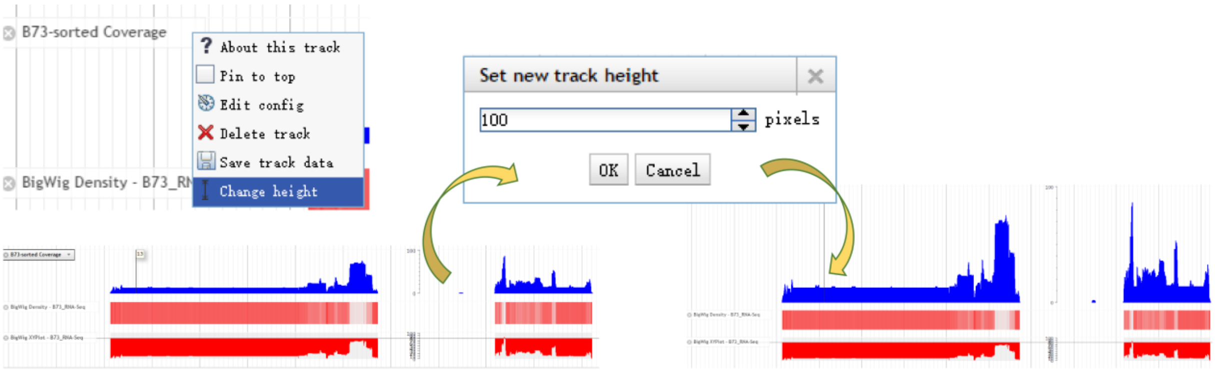
Tracks settings:

All tracks have additional specific settings when click the arrow on the right of track name, the most important one is the “Edit config” and some can be easily modified to better display. Below shows the examples to change height, color or other complex options and related consequences:



```
{
 "maxFeatureSizeForUnderlyingRefSeq": 250000,
 "maxFeatureScreenDensity": 0.5,
 "maxFeatureGlyphExpansion": 500,
 "maxHeight": 200,|
 "histograms": {
   "description": "feature density",
   "min": 0,
   "height": 30,
   "color": "goldenrod",
   "clip_marker_color": "red"
```

```
{
 "maxFeatureSizeForUnderlyingRefSeq": 250000,
 "maxFeatureScreenDensity": 0.5,
 "maxFeatureGlyphExpansion": 500,
 "maxHeight": 600,
 "histograms": {
   "description": "feature density",
   "min": 0,
   "height": 100,
   "color": "goldenrod",
   "clip_marker_color": "red"
```

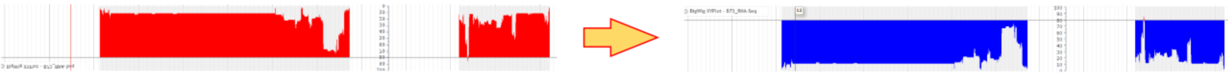


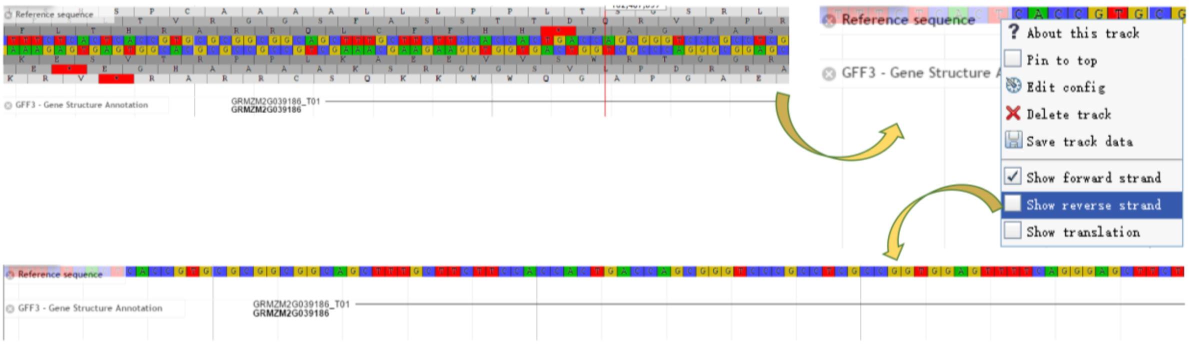
Set new track height

100 pixels

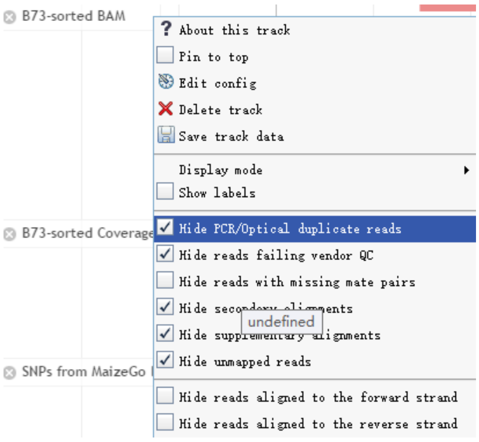
"pos_color": "red",

"neg_color": "blue",





And many settings for reads tracks:



B73-sorted BAM

B73-sorted Coverage

SNPs from MaizeGo

? About this track

□ Pin to top

Edit config

✕ Delete track

Save track data

Display mode ▸

□ Show labels

☑ Hide PCR/Optical duplicate reads

☑ Hide reads failing vendor QC

□ Hide reads with missing mate pairs

☑ Hide secondary alignments

☑ Hide supplementary alignments

☑ Hide unmapped reads

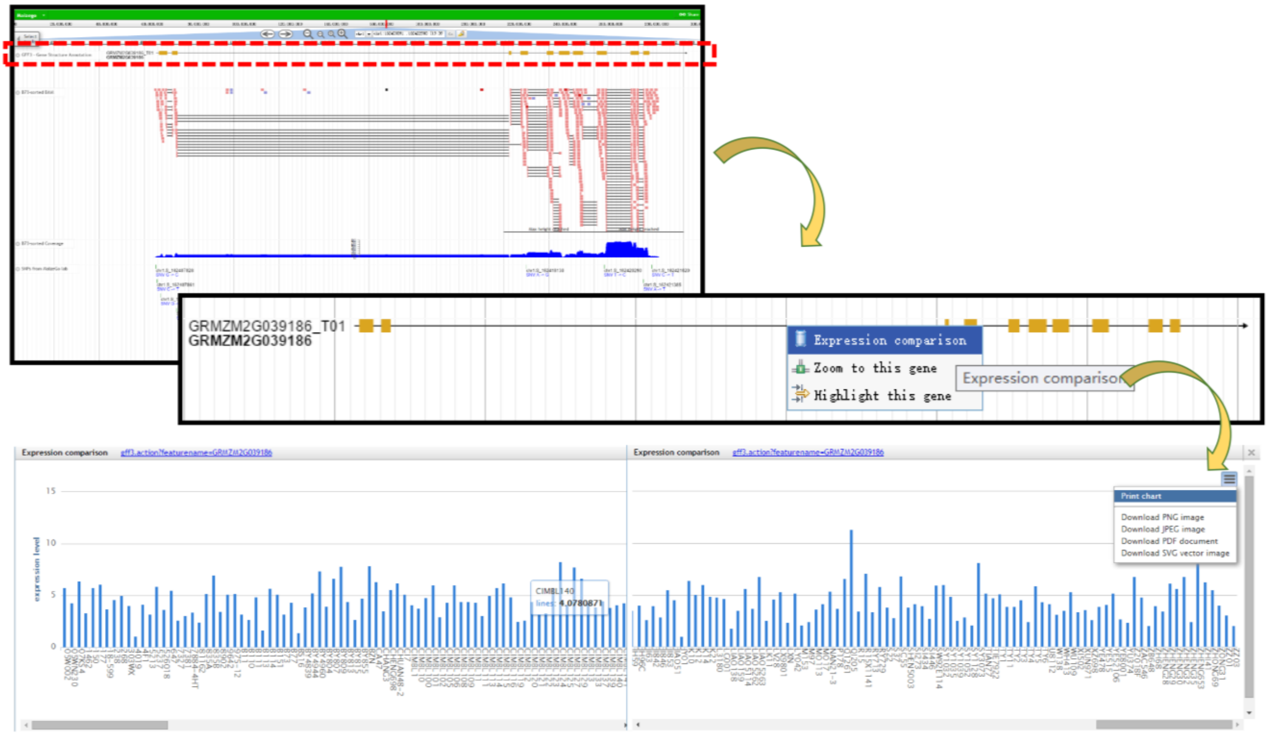
□ Hide reads aligned to the forward strand

□ Hide reads aligned to the reverse strand

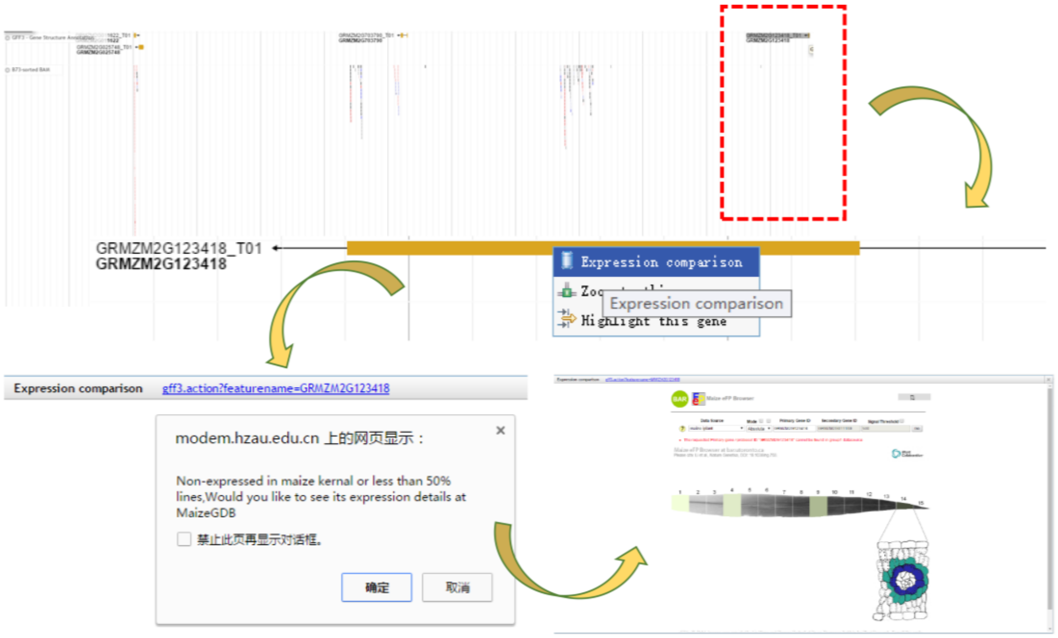
Expression Comparison:

While left-click on the gene body (GFF track) would display the information of specific gene, right-click would give the expression comparison among the whole diverse panel, which takes a easy way to compare the expression quantification with avoiding the large storage and memory in displaying reads coverage of every individual. And if the gene selected expressed less than 50% individuals in our experiment, a re-link would drive users to “maize eFP Browser” to explore the expression tissues of specific candidate.

Here shows an example to compare gene with expressed larger than 50% individuals:



And below shows an example of the gene with expressed less than 50% individuals:

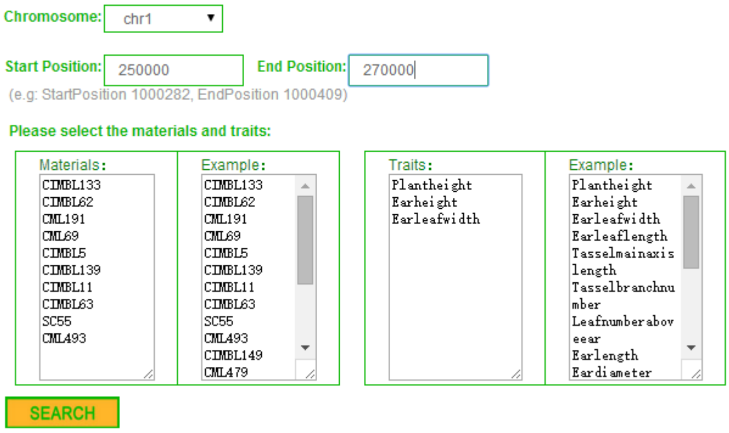


modem.hzau.edu.cn 上的网页显示：

Non-expressed in maize kernal or less than 50% lines,Would you like to see its expression details at MaizeGDB

□ 禁止此页再显示对话框。

Additional TOOLS **(③):** We would provide some tools here that could be used to assist daily researches. Here, we introduce the one developed right now, the simple but useful T-Test analysis tool. We always would like to know which variants within selected region contributed to known phenotype(s), or when we have focused on a specific region, we wonder if this region potentially affect any other traits that we didn't expect before. That expedite tool is proposed to achieve this goal.


Chromosome: chr1 ▾

Start Position: 250000 End Position: 270000

(e.g: StartPosition 1000282, EndPosition 1000409)

Please select the materials and traits:

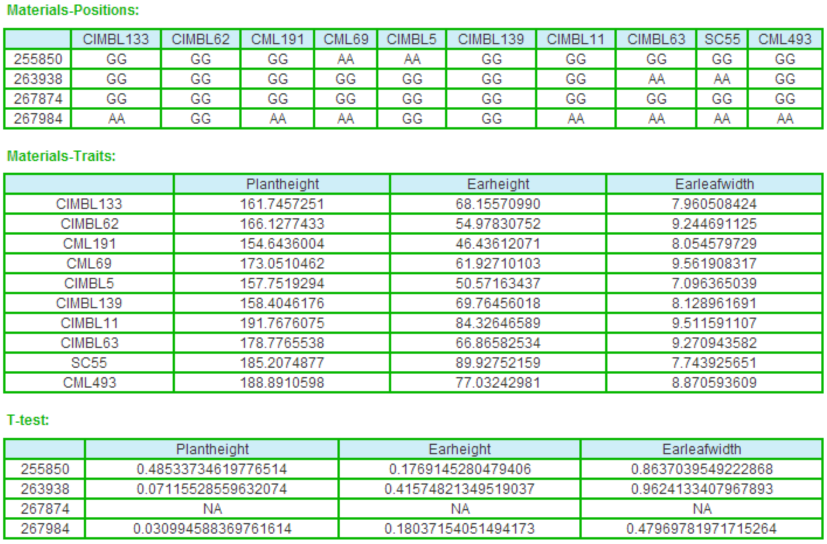
Materials:
CIMBL133
CIMBL62
CML191
CML69
CIMBL5
CIMBL139
CIMBL11
CIMBL63
SC55
CML493

Example:
CIMBL133
CIMBL62
CML191
CML69
CIMBL5
CIMBL139
CIMBL11
CIMBL63
SC55
CML493
CIMBL149
CML479

Traits:
Plantheight
Earheight
Earleafwidth

Example:
Plantheight
Earheight
Earleafwidth
Earleaflength
Tasselmainaxis
length
Tasselbranchnu
mber
Leafnumberabov
eear
Earlength
Eardiameter

Users need to assign a region and group of lines they concerned and hypothetical phenotypes. Based on the different alleles of each SNP, the tool would divide the individuals into two groups and the student T-test is performed to see if the selected phenotype shows difference between the two groups. Above shows a selected region between 250Kb and 270Kb on chromosome 1, and 10 lines with three interested traits. The tools would return these results:


Materials-Positions:

	CIMBL133	CIMBL62	CML191	CML69	CIMBL5	CIMBL139	CIMBL11	CIMBL63	SC55	CML493
255850	GG	GG	GG	AA	AA	GG	GG	GG	GG	GG
263938	GG	GG	GG	GG	GG	GG	GG	AA	AA	GG
267874	GG	GG	GG	GG	GG	GG	GG	GG	GG	GG
267984	AA	GG	AA	AA	GG	GG	AA	AA	AA	AA

Materials-Traits:

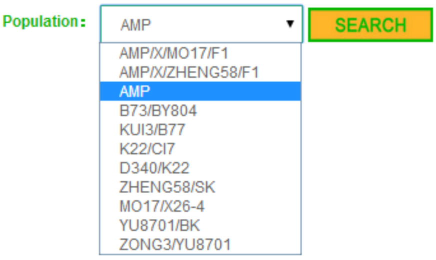
	Plantheight	Earheight	Earleafwidth
CIMBL133	161.7457251	68.15570990	7.960508424
CIMBL62	166.1277433	54.97830752	9.244691125
CML191	154.6436004	46.43612071	8.054579729
CML69	173.0510462	61.92710103	9.561908317
CIMBL5	157.7519294	50.57163437	7.096365039
CIMBL139	158.4046176	69.76456018	8.128961691
CIMBL11	191.7676075	84.32646589	9.511591107
CIMBL63	178.7765538	66.86582534	9.270943582
SC55	185.2074877	89.92752159	7.743925651
CML493	188.8910598	77.03242981	8.870593609

T-test:

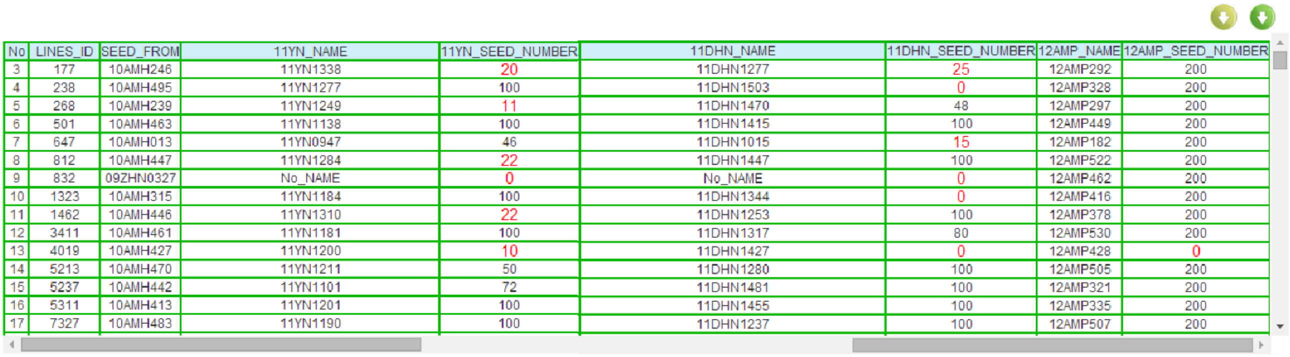
	Plantheight	Earheight	Earleafwidth
255850	0.48533734619776514	0.1769145280479406	0.8637039549222868
263938	0.07115528559632074	0.41574821349519037	0.9624133407967893
267874	NA	NA	NA
267984	0.030994588369761614	0.18037154051494173	0.47969781971715264

The SNP loci within assigned region and alleles of each individual at each site, the trait performance of each individual, and the p-values for every trait at each site. While the "NA" always generated with the reason of no different (or not enough) alleles of those sites within the selected lines. 【**NOTICE:** we highly suggest users to control the selected region with no more than 1M, which may cover too many SNPs and cost too much in computation】

Seed **(④) and *Download* (⑤) management:** The “Seed information” implement mainly help our managers (need to login) to easily administrate the seed storage, delivering and harvest. And external researchers could check if it would be available before submitting their request.



Multiple populations could be selected, and here we take the AMP as example. The search would list the detailed seed information as below, including nomenclature (LINES_ID), source of previous generation for planting (SEED_FROM), and the remaining quantity of each line in each reproduction location (with specific name for corresponding planting plan) is displayed. That seed number below the critical demanded standard would be in red and bold and cannot be loaned any more.



No	LINES_ID	SEED_FROM	11YN_NAME	11YN_SEED_NUMBER	11DHN_NAME	11DHN_SEED_NUMBER	12AMP_NAME	12AMP_SEED_NUMBER
3	177	10AMH246	11YN1338	20	11DHN1277	25	12AMP292	200
4	238	10AMH495	11YN1277	100	11DHN1503	0	12AMP328	200
5	268	10AMH239	11YN1249	11	11DHN1470	48	12AMP297	200
6	501	10AMH463	11YN1138	100	11DHN1415	100	12AMP449	200
7	647	10AMH013	11YN0947	46	11DHN1015	15	12AMP182	200
8	812	10AMH447	11YN1284	22	11DHN1447	100	12AMP522	200
9	832	09ZHN0327	No_NAME	0	No_NAME	0	12AMP462	200
10	1323	10AMH315	11YN1184	100	11DHN1344	0	12AMP416	200
11	1462	10AMH446	11YN1310	22	11DHN1253	100	12AMP378	200
12	3411	10AMH461	11YN1181	100	11DHN1317	80	12AMP530	200
13	4019	10AMH427	11YN1200	10	11DHN1427	0	12AMP428	0
14	5213	10AMH470	11YN1211	50	11DHN1280	100	12AMP505	200
15	5237	10AMH442	11YN1101	72	11DHN1481	100	12AMP321	200
16	5311	10AMH413	11YN1201	100	11DHN1455	100	12AMP335	200
17	7327	10AMH483	11YN1190	100	11DHN1237	100	12AMP507	200

For the administrator, they have more right on the management operations, such as importing, modifying or deleting on the corresponding information. They also need to response to any requirement when received the email on the moment of submission finished. All the management operations can through single or bulk way and each record was kept behind the database and could be directly printed if linking to the printer.

The “Download” management mainly realizes the function to download those multi-omics data in the format of CSV, TXT or XLS. And from here, people could get more information beyond the SEARCH tools offered.

Materials Collection, Data Generation, Processing and Evaluation

Plant germplasm. Currently, we have assembled a global germplasm collection with 527 elite inbred lines (association mapping panel, AMP) released from the major temperate and tropical/subtropical breeding programs of China, CIMMYT and the Germplasm Enhancement of Maize (GEM) project in the US, which were chosen to be representative of maize genetic diversity and/or for their promise in maize improvement. All of the lines were previously assayed by the 50K Maize SNP array. To further explore the genetic mechanisms controlling yield and to increase the yield of maize, deep RNA sequencing was also performed on 368 of the 527 lines using kernels harvested 15 days after pollination (DAP).

Sampling and RNA library construction and sequencing. All 368 lines were planted in two replicate one-row plots in an incompletely randomized block design in Jingzhou, Hubei province of China in 2010. Six to eight ears in each block were self-pollinated, and five immature seeds from three to four ears in each block were collected at 15 DAP. The collected immature seeds from the two replicates were bulked for total RNA extraction, and three additional lines (SK, Han21 and Ye478) were made for biological replicates. Total RNA was extracted using the Bioteke RNA Extraction kit (Bioteke, Beijing, China) according to the manufacturer's protocol and cDNA libraries were constructed according to the manufacturer's standard protocol (Illumina, Inc.). Libraries were amplified by 15 cycles of PCR with Phusion DNA polymerase (New England Biolabs, Inc.) and primers containing barcode sequences to distinguish different libraries during sequencing and data analysis. The average fragment size of each prepared library was 322 bp. Before loading libraries onto the flowcell, the libraries were quantified by qPCR, denatured with sodium hydroxide and diluted to 2.5 pM. Cluster formation, sequencing primer hybridization and 91 cycles of paired-end sequencing were carried out using reagents that Illumina supplied according to the standard protocol.

Reads mapping, SNP calling and quality control. The 368 maize inbred lines were sequenced using 90-bp paired-end Illumina sequencing. After filtering out reads with low sequencing quality, the RNA sequencing (RNA-seq) produced 70.1 million reads for each sample, totalling 25.8 billion high-quality reads. Short Oligonucleotide Alignment Program 2 was used to map the paired-end reads against the B73 reference genome (AGPv2, downloaded from the Maize Genetics and Genomics Database; http://maizegdb.org/). On average, 71.0% of the reads were mapped to the B73 reference genome and 70.3% of the reads mapped to the maize annotated genes (filtered-gene set, release 5b). Among the genes with RNA-seq reads, 71.6% have coverage of >50% of the gene length. Of all the reads mapped to the genome, 83.5% were mapped uniquely and these reads were used to build the consensus sequence for each sample using SAMtools. Briefly, a two-step procedure was used to detect SNPs by carefully considering the characteristics of the RNA-seq data. In the first step, we identified the polymorphic loci in our population, and the population SNP-calling algorithm realSFS was used to calculate the likelihood of variation for each covered nucleotide from the combined data of all of the 368 inbred lines. The variations with probability <0.99 or total depth <50 were filtered out. In the second step, we extracted consensus base, reference base, consensus quality, SNP quality and sequencing depth of each polymorphic locus for each inbred line using the Pileup command, and then considered the consensus base as the individual genotype with the following requirements: if the consensus base was different from the reference base, the non-reference allele must be the same as the non-reference allele detected from the population and the SNP quality must be ≥20. If the consensus base was the same as the reference base, the consensus quality must be equal to or >20 and the minimal depth must be ≥5. For sites that failed to pass these criteria, we regarded the consensus genotype as unreliable and assigned the individual genotype of those sites as missing. The 1,026,244 SNPs with missing rates <0.6 were then left to infer missing genotypes using fastPHASE. Heterozygous genotypes were masked as missing and all SNPs were named according to their physical positions in the reference genome (B73 AGPv2). Altogether, 558,629 SNPs with minor allele frequency (MAF) >5% remained for the subsequent mapping analysis. The concordance rates were >99% between each pair of replicates and 98.6% when compared with overlapping genotypes determined by the MaizeSNP50 BeadChip. Among the 1,026,244 SNPs, 931,484 (90.8%) were mapped to within the 23,106 genes (filtered-gene set, release 5b). On average, there were 40.3 SNPs per gene. Whereas this SNP set includes 69.7% of SNPs reported in a previous study on a nested association mapping population, it contains 7.5 times more exonic SNPs. This not only increases the probability that markers identified possess high linkage disequilibrium with target genes, but also helps in identification of causal variations. Finally, SNPs with a MAF >5% were filtered out and the resulting 525,105 SNPs were then merged with 56,110 SNPs from the MaizeSNP50 BeadChip to produce the merged set of 558,650 SNPs.

Gene expression profiles. To quantify the expression of known genes, reads that uniquely mapped to each gene within the reference genome (filtered-gene set, B73 AGPv2) were summed and normalized according to RPKM (reads per kilobase of exon model per million mapped reads). On average, there were 1540.7 reads for each whole gene for each individual. Genes with a median expression level (in RPKM) larger than 0 (38,850), and having mapped sequencing reads (expressed) in more than half of the maize lines (28,769) were used for further eQTL mapping. The expression values of each gene were then normalized using a normal quantile transformation to meet the assumption of detecting eQTLs through a linear mixed model that the expression values follow a normal distribution. This quantile transformation does not fully solve the problem, however, it is a simple, sensible way to guard against strong departures from modelling assumptions with the small effect sizes typical in genetic association studies.

Maize kernel metabolome profiling, identification and annotation. To extract metabolites of maize kernels, the association panel lines were planted in one-row plots in an incompletely randomized block design at three locations in China: Hainan (Sanya, E 109º51', N 18º25') in 2010 and Yunnan (Kunming, E 102º30', N 24º25') and Chongqing (E 106º50', N 29º25') in 2011. All inbred lines were self-pollinated and ears of each plot were hand-harvested at their respective physiological maturity, followed by air drying and shelling. For each line, ears from five plants were harvested at the same maturity and 12-well growth kernels were randomly selected from five plants and bulked for grinding by using a mixer mill (MM 400, Retsch) with zirconia beads for 2.0 min at 30 Hz. The powder of each genotype was partitioned into two sample sets and stored at −80ºC until extraction. One sample set was extracted for lipid-soluble metabolites, while the other was extracting for water-soluble metabolites. One hundred mg of powder and 1 ml absolute methanol, which contained 0.1 mg/l each of lincomycin and lidocaine, were used for lipid-soluble metabolites (or 70% methanol for water-soluble metabolites). Samples were extracted overnight at 4ºC. After centrifugation at 10,000 g for 10 min, 0.4 ml of each extract was combined and filter spun using 0.22-μm filters (ANPEL, Shanghai, China, http://www.anpel.com.cn/) before analysis using an LC-ESI-MS/MS system. The metabolite quantification and annotation was performed by our newly developed method. To facilitate the identification/annotation of detected metabolites by our widely targeted metabolomics approach, accurate m/z of each Q1 was obtained, if possible. To this end, extracted ion chromatograms of the ESI-QqTOF-MS data for each of Q1 (m/z ± 0.2 Da) of the 983 transitions in the MS/MS library were manually evaluated for the presence of the target substances by analysing corresponding mass spectra, and accurate m/z values were obtained. For each of the corresponding accurate m/z, a fragmentation pattern was obtained by running the analysis under targeted MS/MS mode using three different collision energies of 10, 20 and 30 eV. The accurate m/z was assigned to the corresponding Q1 if similar fragmentation patterns were obtained between the ESI-Q TRAP-MS/MS and the ESI-QqTOF-MS/MS. Eventually, an accurate mass of 245 of Q1 was obtained. The MS/MS library was annotated based on the fragmentation pattern (delivered by ESI-Q TRAP-MS/MS and/or the accurate m/z value delivered by ESI-QqTOFMS/MS) and the retention time of each metabolite. Based on the annotation, commercially available standards were purchased and analysed using the same profiling procedure as the extracts. By comparing the m/z values, the retention time and the fragmentation patterns with the standards, 49 metabolites were identified, including amino acids, flavonoids and fatty acids (such as a-linolenic acid), and some phytohormones. For the metabolites that could not be identified by available standards, peaks in the

MS/MS library, especially the peaks having similar fragmentation patterns with the metabolites identified by authentic standards, were used to query the MS/MS spectral data taken from the literature or to search the databases (MassBank, KNApSAcK, HMDB, MoTo DB and METLIN). Best matches were then searched in the Dictionary of Natural products and Kyoto Encyclopaedia of Genes and Genomes for possible structures. In all, 184 metabolites were identified and more than four different pathways were detected.

Population structure, relatedness matrix and association analysis. A subset of 16,338 SNPs with <20% missing data and MAF >5% were used to estimate population structure and kinship coefficients. STRUCTURE was used to infer population structure with 10,000 replications for burn-in and MCMC processes, and five runs were performed at k = 3. The samples were divided into three subgroups, as previously suggested for this panel. The associations between the extracted SNPs with MAF>5% and all measured traits (including morphological phenotypes, kernel metabolome and transformed expression traits) were analysed using the linear mixed model (MLM) incorporating population structure and kinship using TASSEL. The significance cut off was set generally to 1/N, where N is the number of markers used, and all the results were provided to be displayed. The top six hidden confounding factors determined to be contributing to expression variability by Bayesian factor analysis (implemented in PEER) were additionally included in the mixed model, in addition to population structure, to examine the validity of association significance for eQTL mapping.