Shine a Light on MAGIC Data

Brian S Yandell, http://www.stat.wisc.edu/~yandell

Jackson Labs, 18 October 2016

shine a light on MAGIC data

Shiny tool to study DO and other MAGIC crosses

- What is MAGIC?
- Allele vs SNP Scans
- Additive Model is Quick & Easy
- Gene/SNP Action
- Shiny under the Hood

what is MAGIC?

- Multiparent Advanced Generation Inter-Cross (MAGIC)
  - experimental populations with $>$2 segregating alleles
- advanced generations yield many meiotic events
  - typically low linkage disequilibrium
  - capable of fine mapping in one pass
- de Koning DJ, McIntyre LM (G3 2014)
  - animals: mouse, Drosophila
  - plants: maize, wheat, rice, sorghum, Arabidopsis, Pigeonpea
  - mapping populations: AIL, CC, HS, DO, DSPR

Attie/Jax DO population

- 8 CC founder strains (generation 19-21)
- 400 mice in 4 waves
- multiple traits measured
  - clinical traits (insulin secretion)
  - 150K SNPs, 30K RNA-Seq
  - proteomic, metabolomic, lipidomic

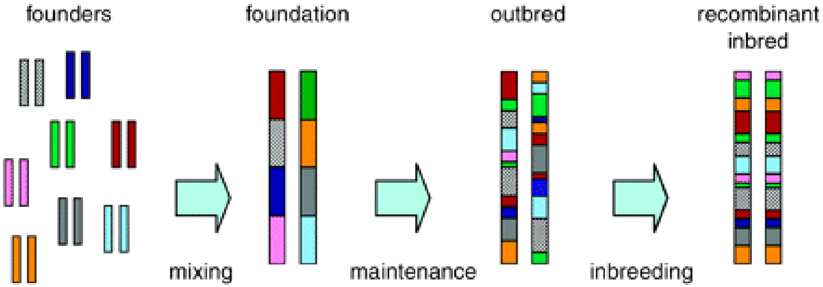


Figure 1: Valdar et al. 2006 doi:10.1534/genetics.104.039313

Figure 2: http://compgen.unc.edu

allele vs SNP scans

- allele-based genome scan: LOD maps
  - continuous curve across loci
  - interval mapping for missing data
  - model effect of founder alleles
- DO founder alleles: A,B,C,D,E,F,G,H
- response $\sim$ sum of effects of alleles
- additive model
  - $y \sim a_1 + a_2$ (1st & 2nd allele)
  - test if all a's are the same
  - 8 unknown parameters

simple story on chr 1

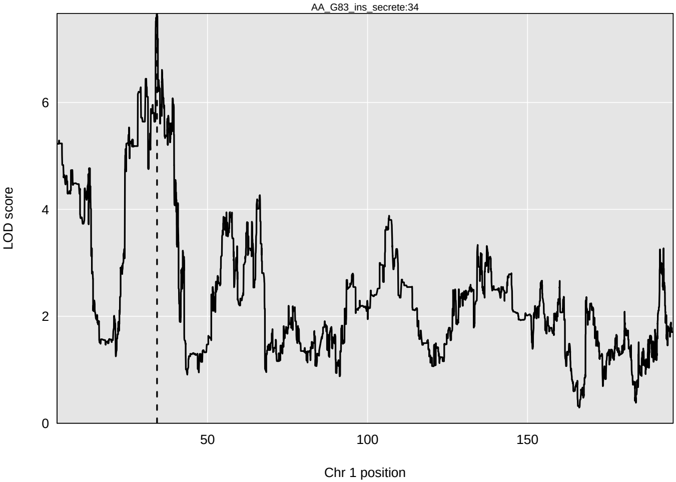


Figure 3: chr 1 scan

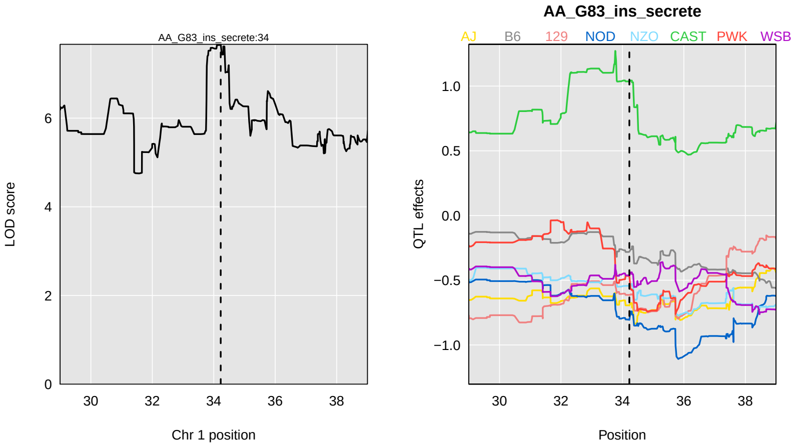


Figure 4: chr 1 zoom scan & effects

SNP-based genome scan

- SNP-based genome scan: GWAS Manhattan plots
  - discrete tests of SNPs or other features
  - typically 2 SNP alleles (1 is reference)
  - model effect of number of non-ref SNP copies
- SNP recorded as pair of DNA base pairs (A,C,G,T)
  - SNPs typically have two values (G/T)
  - individual has genotype GG, GT or TT
- simplifed to number of copies of non-reference allele
  - $s = 0,1,2$
  - DO reference is B = B6
- additive model:
  - $y \sim a + bs$ (2 unknowns)
  - test slope: $b = 0$?

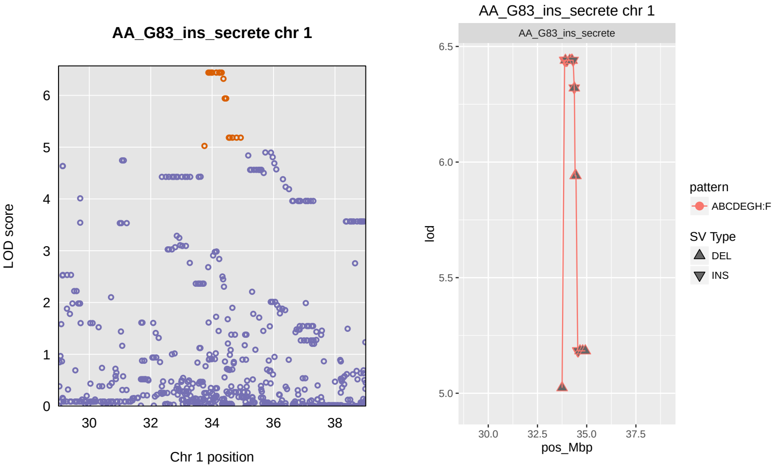


Figure 5: SNP scan & top pattern

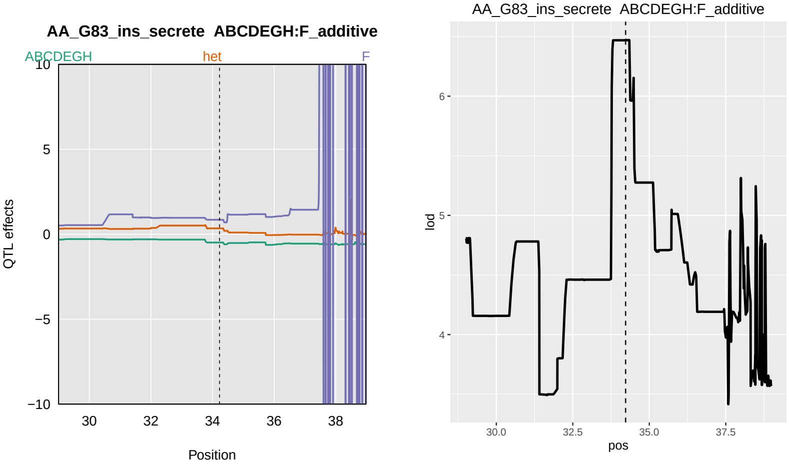


Figure 6: LOD & allele contrast scan

additive model is quick & easy

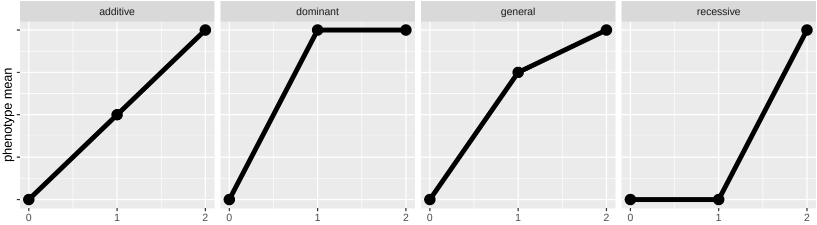
- good news
  - fewer parameters to understand
  - easy to build tools
  - fast to run
  - relate LOD to allele effects directly
  - likely good model for mRNA expression
- bad news
  - could miss important gene action information
  - could miss loci (false negatives)
  - could detect false positives
  - LOD and allele effects may conflict due to dominance

full model (with dominance)

- DO founder alleles: A,B,C,D,E,F,G,H
- response $\sim$ effect of pair of alleles
- full model for allele-based scan
  - $y \sim a_1 + a_2 + d_{12}$ (additive & dominance effects)
  - $y \sim \texttt{mean}(A_1, A_2)$ (alleles A_1, A_2)
  - test if all means are the same (all *a*s equal, all *d*s zero)
  - 36 unknown parameters
- full model for SNP-based scan
  - $y \sim \texttt{mean}(s)$
  - test if all means are the same
  - 3 unknowns

gene/SNP action

- study additive & dominance of single trait
- compare co-mapping traits – same gene action?
- extend discrete SNP-based scan to continuous
  - scan of particular allele contrasts
  - LOD and allele contrast scans



complicated story on chr 3

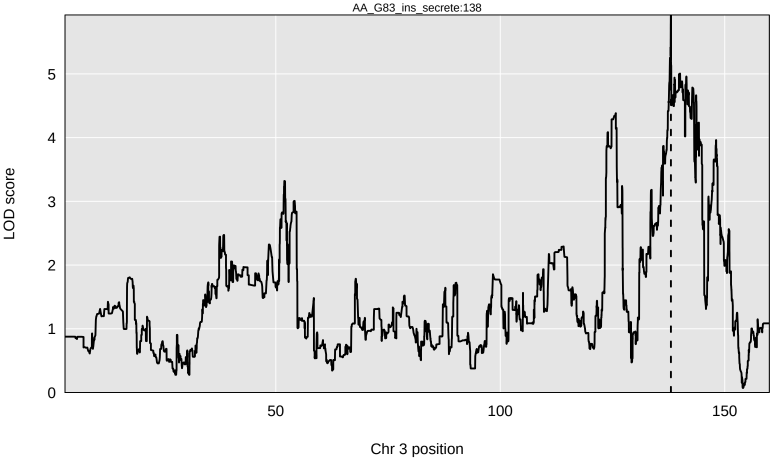


Figure 7: chr 3 scan

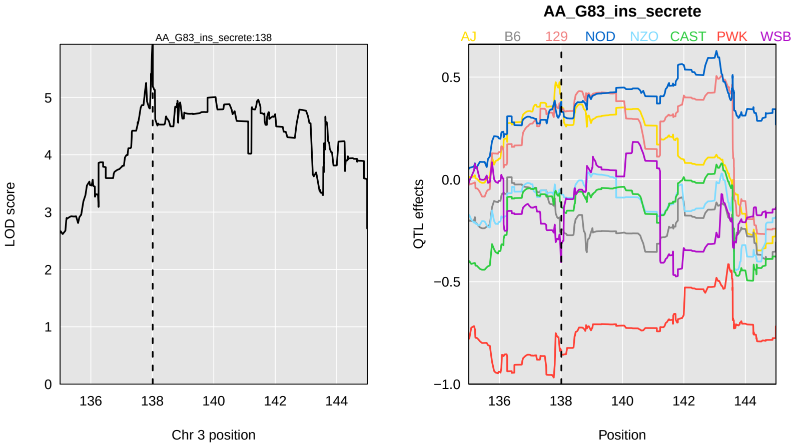


Figure 8: chr 3 zoom scan & effects

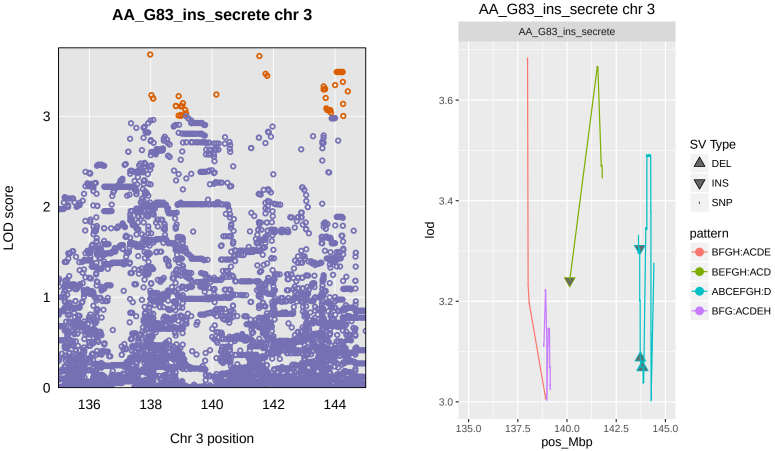


Figure 9: SNP scan & top pattern

additive & dominance together

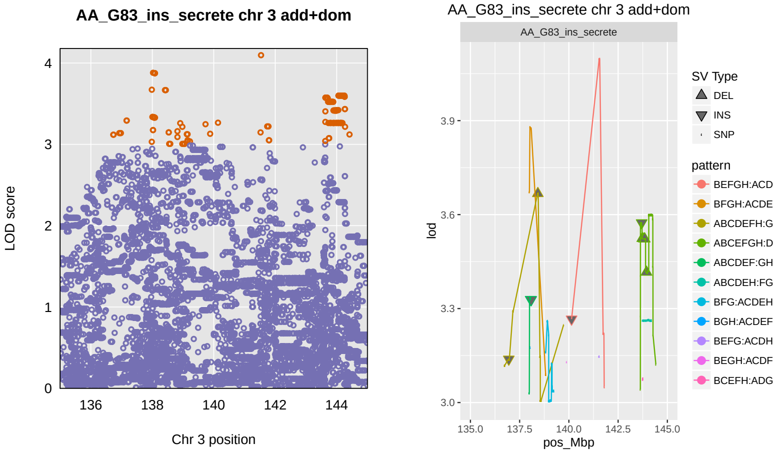


Figure 10: 3-level SNP scan & top pattern

general SNP contrasts

- identify 3-level SNP with strong LOD
- interpret as contrast involving 36 allele pairs
  - example AB:CDEFGH
- divide subjects into 3 groups at each locus
  - AA,AB,BB allele pairs (3)
  - CC,CD,. . . ,GH,HH allele pairs (21)
  - rest allele from AB and allele from CDEFGH (12)
- scan region for this set of contrasts
  - LOD scan with 2 df of AB:het:CDEFGH
  - allele group scan

additive BFGH:ACDE

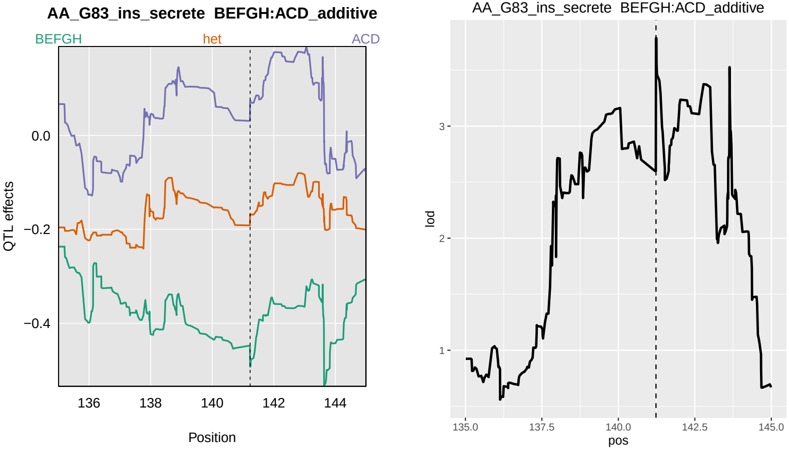


Figure 11:

B=B6 recessive, D=NOD dominant

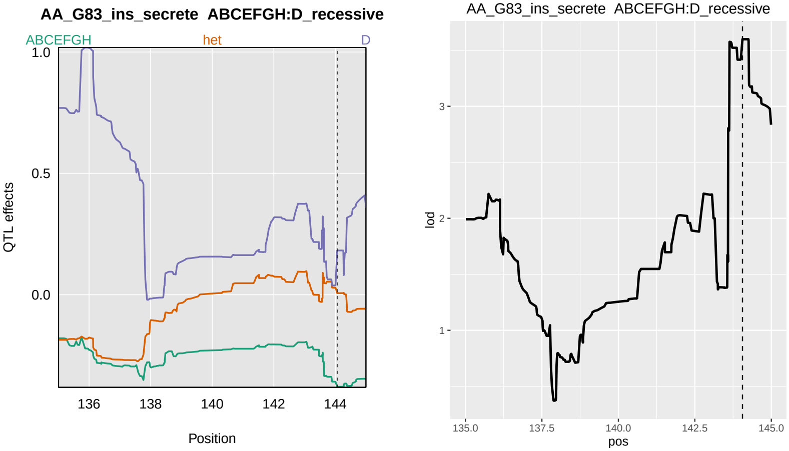


Figure 12:

B=B6 dominant, G=PWK recessive

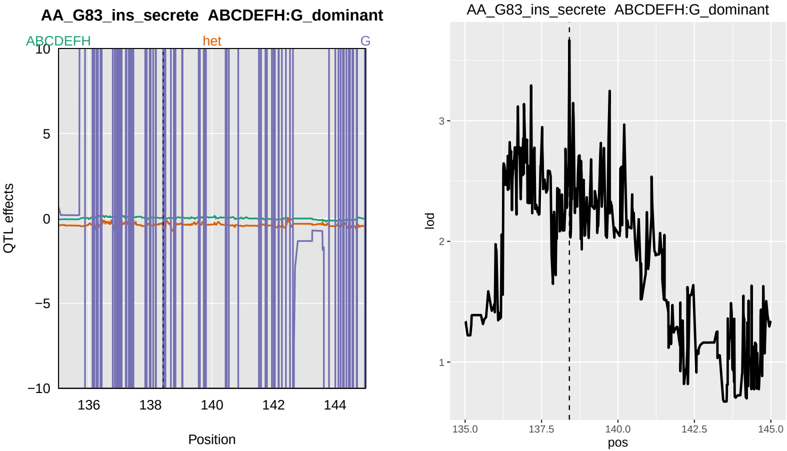


Figure 13: unstable: few double recessive GG

shiny under the hood

- work flow
- tools and resources
- challenges remaining

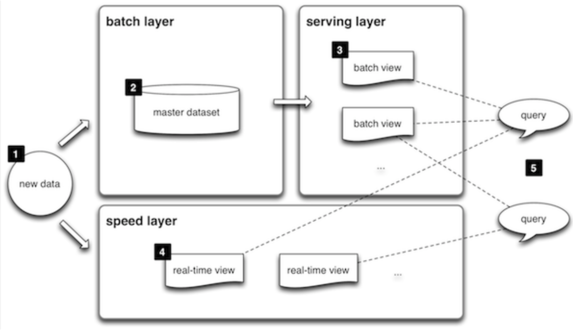


Figure 14: lambda architecture (Nathan Marz, Twitter)

lambda architecture for doqtl2

- batch layer (DOQTL)
    - genotypes
    - phenotype LOD scans across genome
    - phenotype LOD & allele scans for chromosome
    - phenotype SNPs & genes in peak region
- serving layer
    - web serving layer
- speed layer
    - Shiny "speed" layer
    - phenotype LOD & allele scan for chromosome
    - detailed look at peak region

tools & resources

- Original batch pipeline: DOQTL (Dan Gatti et al. 2014 G3)
- R/qtl2 (Karl Broman) (Karl Broman et al. in progress)
    - qtl2geno, qtl2scan, qtl2plot
    - web site
- new R packages
    - R/qtl2 package suite (Karl Broman)
    - R/doqtl2 package
    - R/qtl2shiny package
- Derived Data in various file forms
    - SQLite for mouse SNPs and gene features
    - RDS R data objects for genotypes
    - CSV comma separated variable for phenotypes

challenges remaining

- Shiny user interface from Rstudio
  - translates from R to HTML via javascript
  - run under Rstudio or via Shiny server
  - R/qtl2shiny has 20+ shiny modules
  - translate to Python?
- analysis & computation
  - calibration of significance thresholds
  - dimension reduction (filtering) for massive phenotypes
  - incorporation of causal network tools
- operation & connectivity
  - connecting to other omic resources
  - fast management of raw, derived and intermediate data
- visualizations
  - scalable & interactive (D3) visual displays
  - automated report generation

data storage issues

- raw data
- derived data
  - 3Gb gene, SNP & SVS features (mouse)
  - 2Gb haplotype probabilities
  - 10Gb diplotype probabilities
  - 0.2Mb clinical phenotypes
  - 1-2Gb molecular phenotypes
  - spatial/image phenotypes?
- intermediate data
  - tables & summaries
  - plot data (or saved plots)
  - on-the-fly vs pre-stored vs as-needed
- portability issues
  - CSV vs RDS vs SQL vs . . .
  - common pool vs on-site

software issues

- R analysis & visualization
  - QTL tools: qtl2 squite: geno, scan, plot
  - discovery: doqtl2, qtl2shiny
  - interactive: shiny, shinydashboard
  - wrangle data: dplyr, tidyr, readr, stringr
  - graphics: ggplot2, grid
- phenotype & genotype pipelines
  - variety of languages and formats
  - in domain of biologists (or chemists or . . .)
- high volume pipelines
  - hadoop/map-reduce technology
  - high throughput phenotypes
  - data resampling for thresholds
- scaling up to muli-user interactive system