plant systems genetics:

from markers to whole genomes

Brian S. Yandell, UW-Madison, www.stat.wisc.edu/~yandell/PlantSysGen

January 2017

outline

- Introduction [PDF | HTML] (32 pages)
- Quantitative Trait Loci (QTL) [PDF | HTML] (43)
- Association Mapping [PDF | HTML] (24)
- Genome-Wide Selection [PDF | HTML] (12)
- Multiple Traits [PDF | HTML] (18)
- Systems Genetics Tools [PDF | HTML] (14)

http://www.stat.wisc.edu/~yandell/talk/PlantSysGen

overview

Systems genetics is an approach to understand the flow of biological information that underlies complex traits.

how to relate phenotype to genotype

- genetic effects (QTL & polygenes)
- prediction & selection (MAS, GS)

with changing technology

- laboratory protocols
- statistical methods
- computational tools

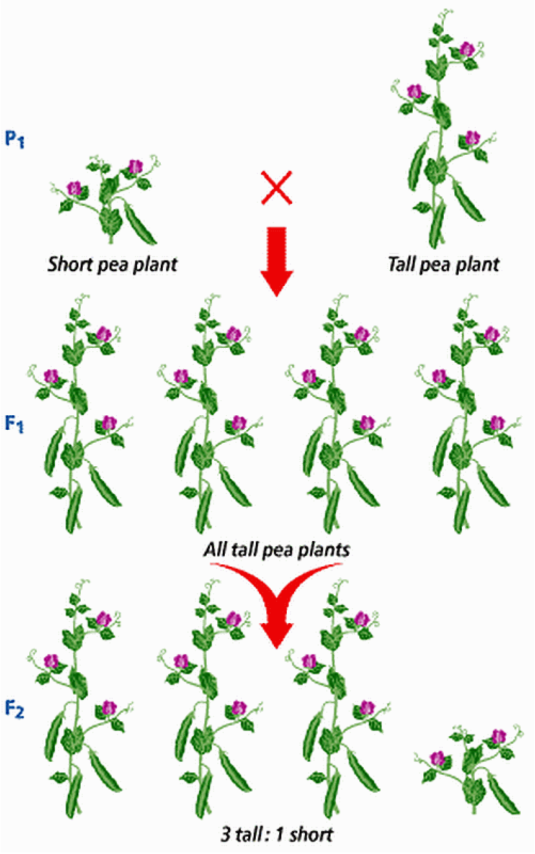


Figure: plantcellbiology.masters.grkraj.org

goal of system genetics studies

- predict performance of future offspring
    - genome-wide selection
- estimate genetic architecture of traits
    - quantitative trait loci (QTL)

Great time to become involved in modern approaches!

- many challenges
- many opportunities for substantial contributions
- help unravel important problems in biological systems
- data tools are maturing

does genotype influence phenotype?

Goals for genetic architecture

- identify quantitative trait loci (QTL)
    - (and interactions among QTL)
- find interval estimates of QTL location
- estimate QTL effects

Goals for predicting future performance

- predict breeding value of individuals
- select best individuals using genome

PHE = GEN + ENV

phenotype = genotype + environment

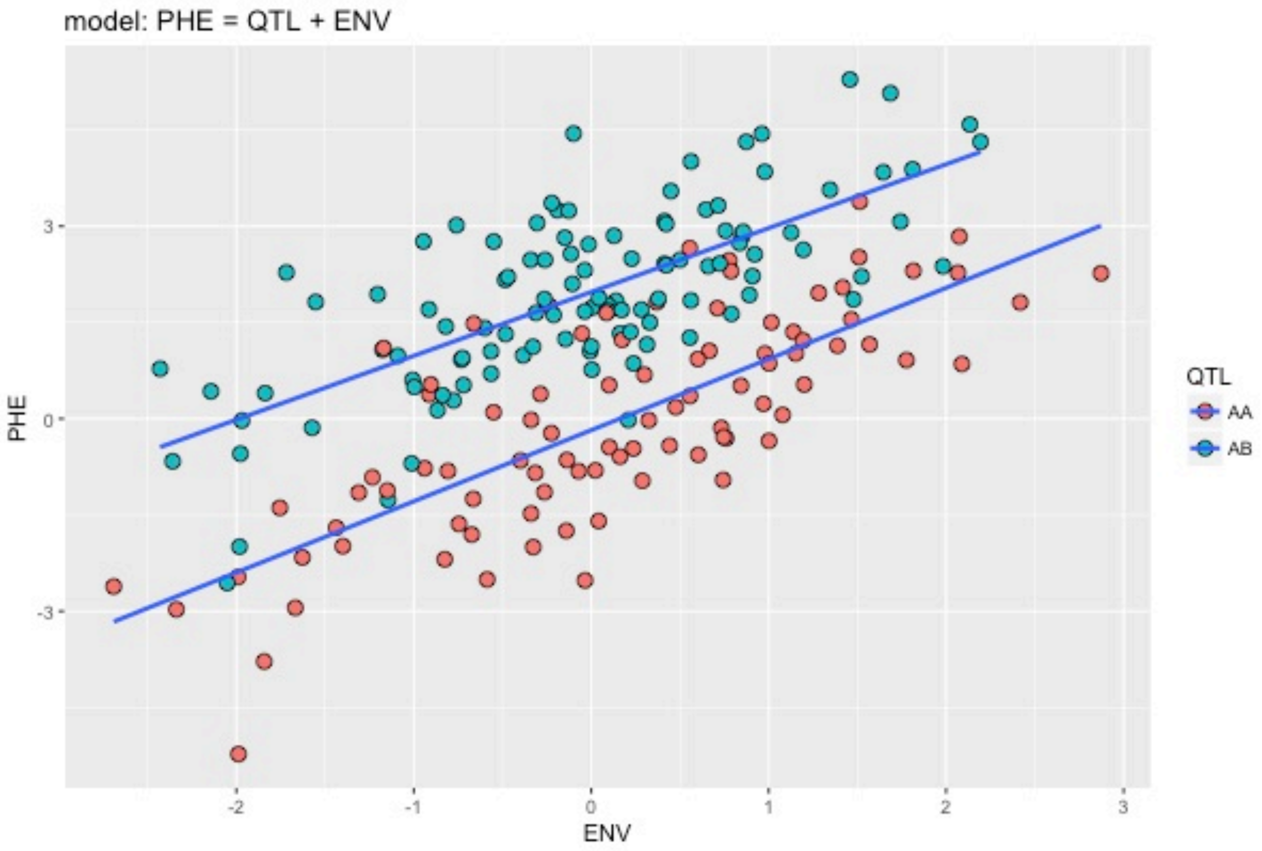
- GEN = QTL + poly
    - genotype = local + polygenic effects
- ENV = design + predictors + error
    - design factors (blocks, locations, …)
    - predictor variables (heat, light, soil additives, …)
    - measurement error (independent)

Falconer & Mackay (1960–1996)

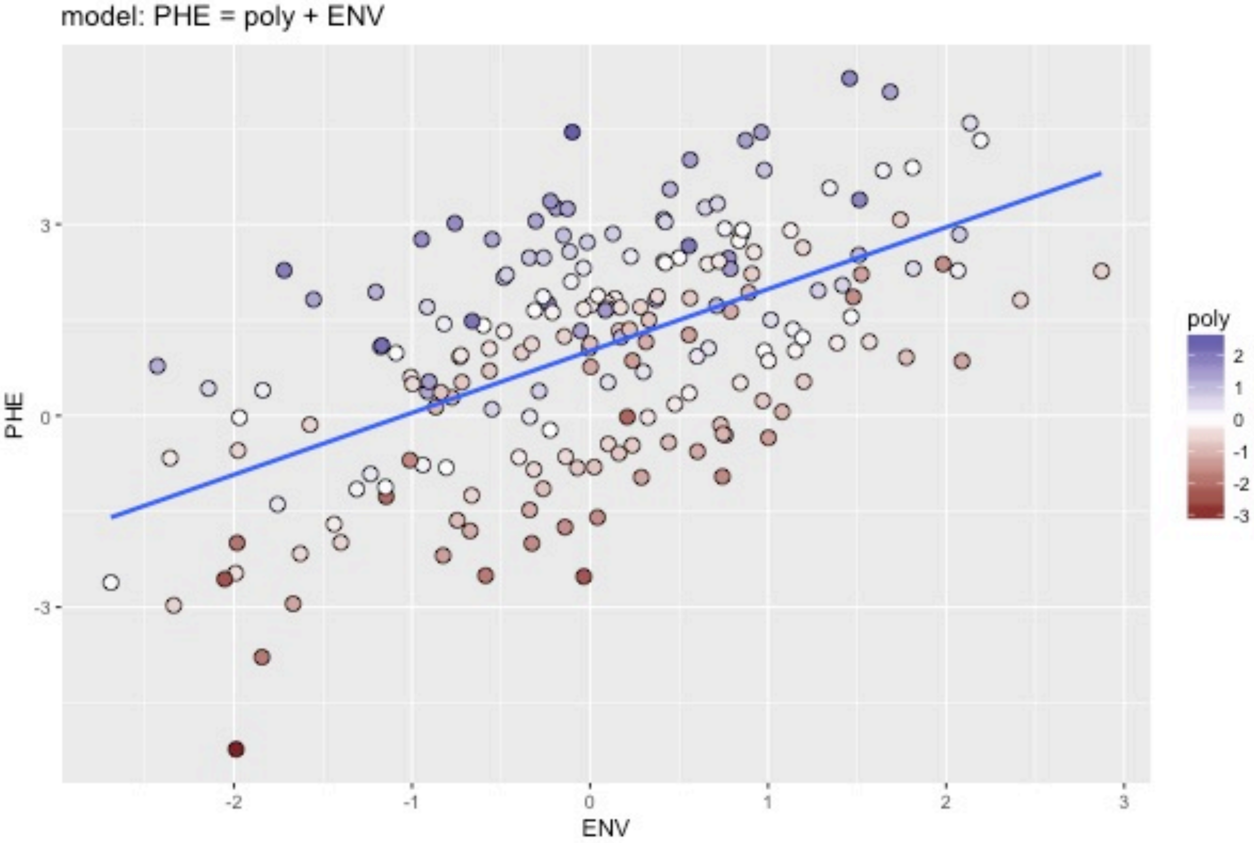
GEN = QTL + poly

- QTL: quantitative trait loci
    - local to an identified genomic region
    - large Mendelian effects on mean
- poly: polygenic association across genome
    - depends on population structure (kinship)
    - measures relationships away from QTL
    - average of many small effects

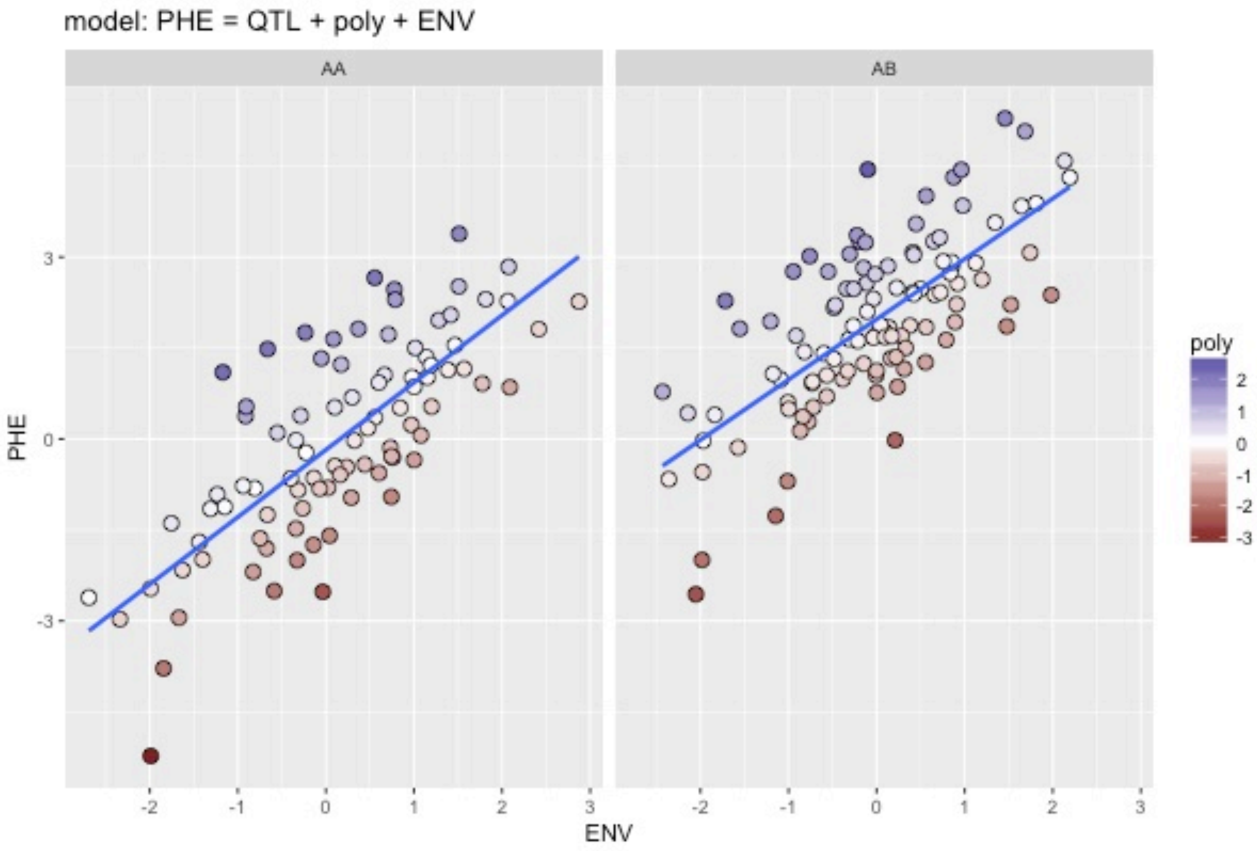
PHE = GEN + ENV example



PHE = GEN + ENV example



PHE = GEN + ENV example



thanks up front

- Karl Broman, UW-Madison
- Jeff Endelman, UW-Madison
- Guilherme Rosa, UW-Madison
- Eleazar Eskin, UCLA
- Gary Churchill, Jackson Labs
- Alan Attie, UW-Madison
- UW-Madison sabbatical program
- Kasetsart Univeristy, Thailand
  (Piya Kittipadakul & Janejira Duangjit)

Olbrich Botanical Garden, Madison, WI

UW-Madison collaboration

- Plant Breeding & Plant Genetics Program
    - Statistical Genetics & Genomics Focus
- Biometry Program
- Biostatistics & Medical Informatics Department (BMI)
- Statistics Department
- Laboratory of Genetics
- Animal Breeding & Genetics

UW-Madison Biometry Program

- joint faculty with Statistics
  - Cecile Ane (Botany)
  - Murray Clayton (Plant Pathology)
  - Brian Yandell (Horticulture)
  - Jun Zhu (Entomology)
- collaborative research & consulting
- teaching courses at all levels
  - introductory data science methods
  - Bayesian methods, spatial statistics
- Biometry Masters program

UW Biometry Consulting model

- faculty & staff time paid by UW (CALS)
  - no visit cost
  - builds long-term collaboration
  - not limited by program/project size
- mentoring of research enterprise
  - gradaute student training
  - faculty & staff relationship building
  - encourage research teams for grants
- campus-level vision of data & research
  - and human capital

RA Fisher (1948) defined biometry

Biometry is “the active pursuit of biological knowledge by quantitative methods … [through] constant experience in analysing and interpreting observational data of the most diverse types…. [W]e come to think of ourselves … in terms of the community of our interests with those doing similar work in other departments."

at inaugural meeting of the Biometric Society

UW-Madison Biostat & Med Info (BMI)

- faculy expertise in variety of research areas
- collaborations large in human health
  - but extended across campus
- statistical genetics & genomics
  - Newton, Kendziorski, Keles, Dewey, Broman, Wang
- bioinformatics
  - Shavlik, Page, Craven, Dewey, Coen, Roy, Gitter
- image analysis
  - Dyer, Chung, Singh
- affiliate faculty
  - Gianola, Rosa, Yandell

UW-Madison community

- plants (Endelman, de Leon Gatti, Guttierez)
- animals (Rosa, Gianola, Kirkpatrick)
- genetics (Payseur, Doebley)
- microbes (Gasch, Rey, …)
- evolution & phylogenetics (Ane, Larget, Baum, Spooner)
- high throughput methods
    - computers: Livny, Negrut, Wilson
    - chemistry: Coon, Pagliarini
    - botany: Spalding

approach in these talks

mix of presentation style to plant-based audience

- theory
  - set the stage
  - show big picture
- applied: using R packages
  - qtl: basic gene mapping
  - rrBLUP: genome-wide prediction & polygenes
  - qtl2: high throughput gene mapping
- source: https://github.com/byandell/PlantSysGen
- slides: http://www.stat.wisc.edu/~yandell/talk/PlantSysGen

challenges in systems genetics

simpler models yield clearer results

- compare 2 conditions
- examine linear trend
- control for other factors

but reality may be more complicated

- masking of genetic effect (by background, etc.)
- subtle timing (when to measure)
- hard to measure key features (shape, quality)
- unknown details of processes under study

evolution of laboratory protocol

genetic information (genotype)

- genetic markers discovered by accident (RFLP,…)
- dense sets of polymorphic markers (SNP, GBS)
- whole genomes sequencing

trait information (phenotype)

- physiology (internal) & environment (external)
- molecules & images
- inexpensive, high volume assays $100 - 10,000s$ of plants

(individual cell technologies not covered here)

genotyping

- RFLPs & other early technologies
- structural variants
    - SNPs (single nucleotide polymorphisms)
    - InDels, inversions, larger blocks (100s-1000s of bps)
    - huge blocks (20K+ bps)
- GBS (genotype by sequence)
- read genotype from RNA-Seq
- Cautions:
    - missing data, mistakes in reads, sample mixups
    - biases in technologies
    - reference sequence vs other founders

evolution of statistical methods

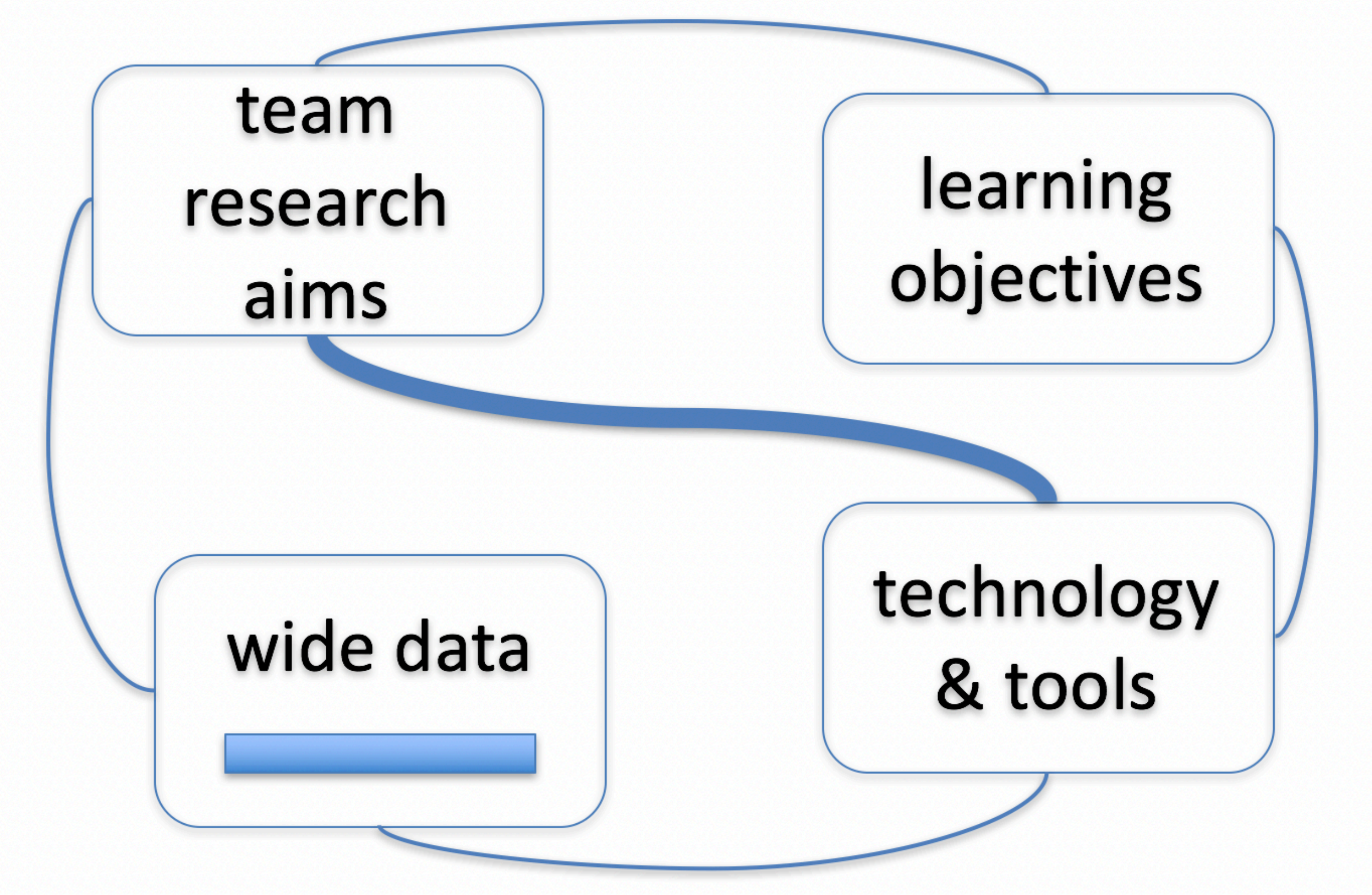
- experimental design: how populations are created
    - two-founder experiments (backcross, intercross)
    - advanced crosses (RILs)
    - multi-parent populations (MPP)
- model selection: how phenotypes relate to genotypes
    - single marker regressions & interval mapping (QTL)
    - association mapping (including polygenes)
- estimation and prediction
    - genetic action (additive, dominance, epistasis)
    - marker assisted (MAS) & genomic selection (GS)

evolution of computational tools

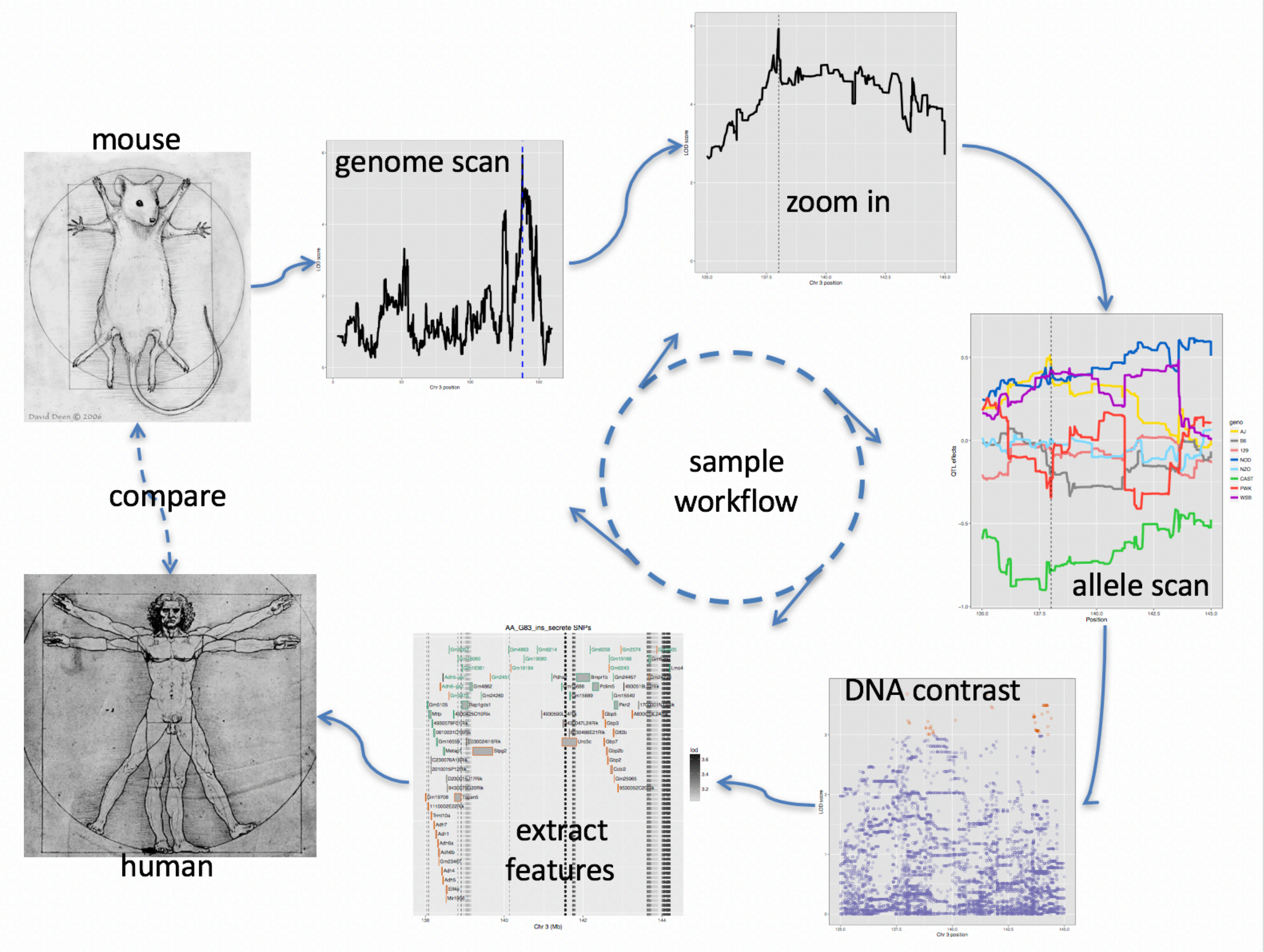
Advances in measurement, design and analysis would be academic without advances in computational technology.

- faster machines -> faster throughput of more stuff
- methods translated into algorithms
  - open source code: freely distrubuted, easy to study
  - standalone programs
  - packages in language systems (R or Python or Matlab)
- collaboration and sharing
  - interconnectivity of algorithms and data resources
  - collaboration tools – beyond email attachments
  - emerging collaboration systems

tools & workflow: big idea



team research aims



mouse

genome scan

zoom in

allele scan

DNA contrast

extract features

human

compare

sample workflow

communication challenges

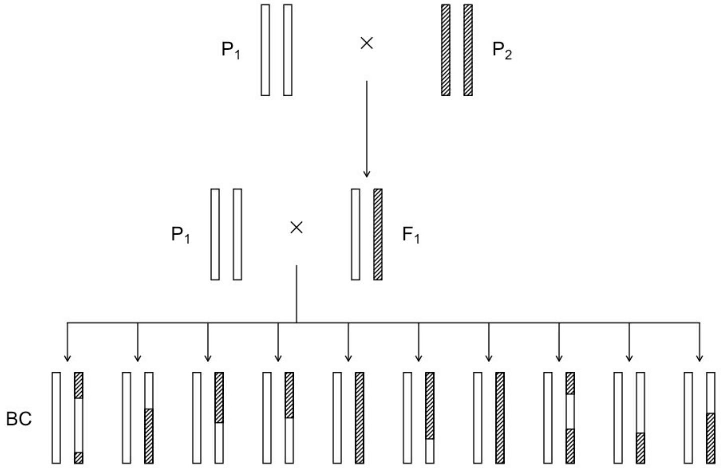
- English as 2nd, 3rd (4th?) language
- data experience and learned patterns
- stat experience and access to consultants
- math anxiety (see Sheila Tobias books)
- IT/computing experience and access to tools
- genetics knowledge
- communicating outside chosen field

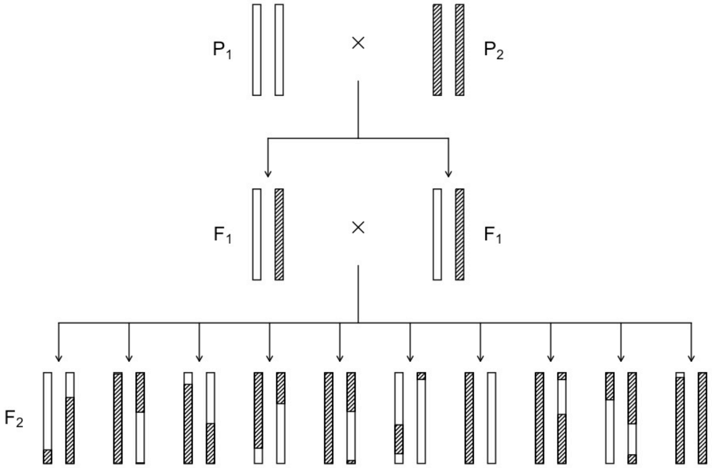
Experimental Designs

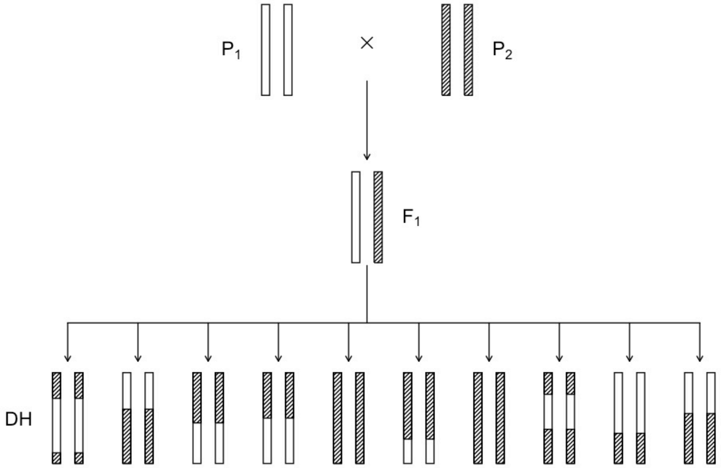
- common breeding designs
  - backcross (BC)
  - intercross (F2)
  - doubled haploid (DH)
- advanced intercross lines
  - recombinant inbred lines (RILs)
  - near isogenic lines (NILs) & consomics
  - multi-parent populations (MPP)

common breeding designs

- 2 (inbred) founder alleles
- 2 generations
- backcross (BC): 1 meiosis
- doubled haploid (DH): 1 meiosis
- intercross (F2): 2 meioses



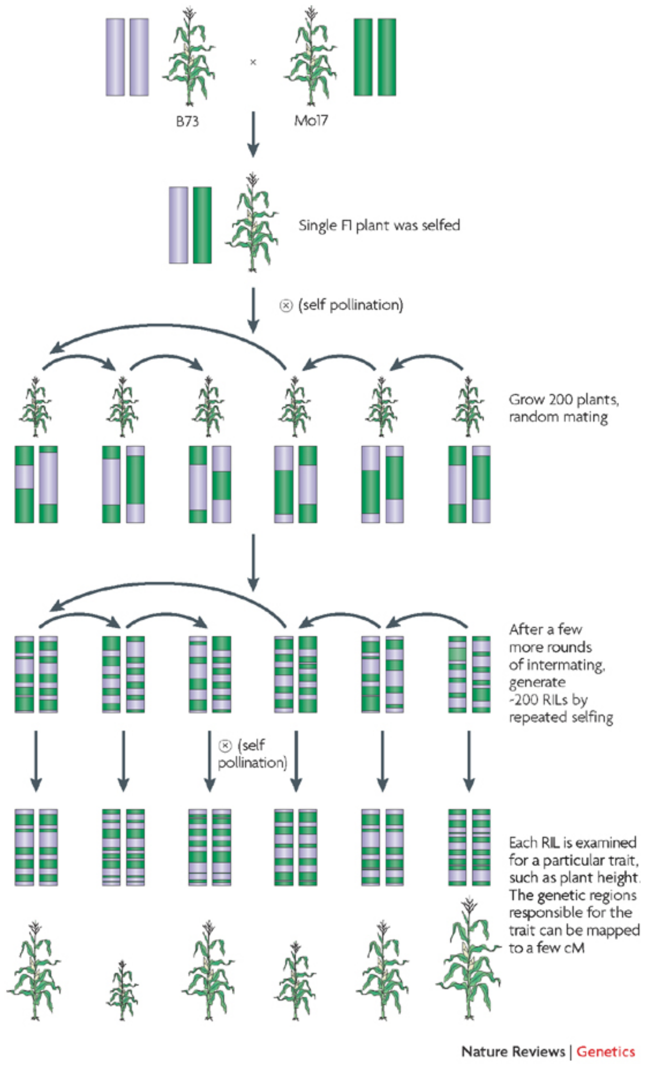




recombinant inbred lines (RIL)

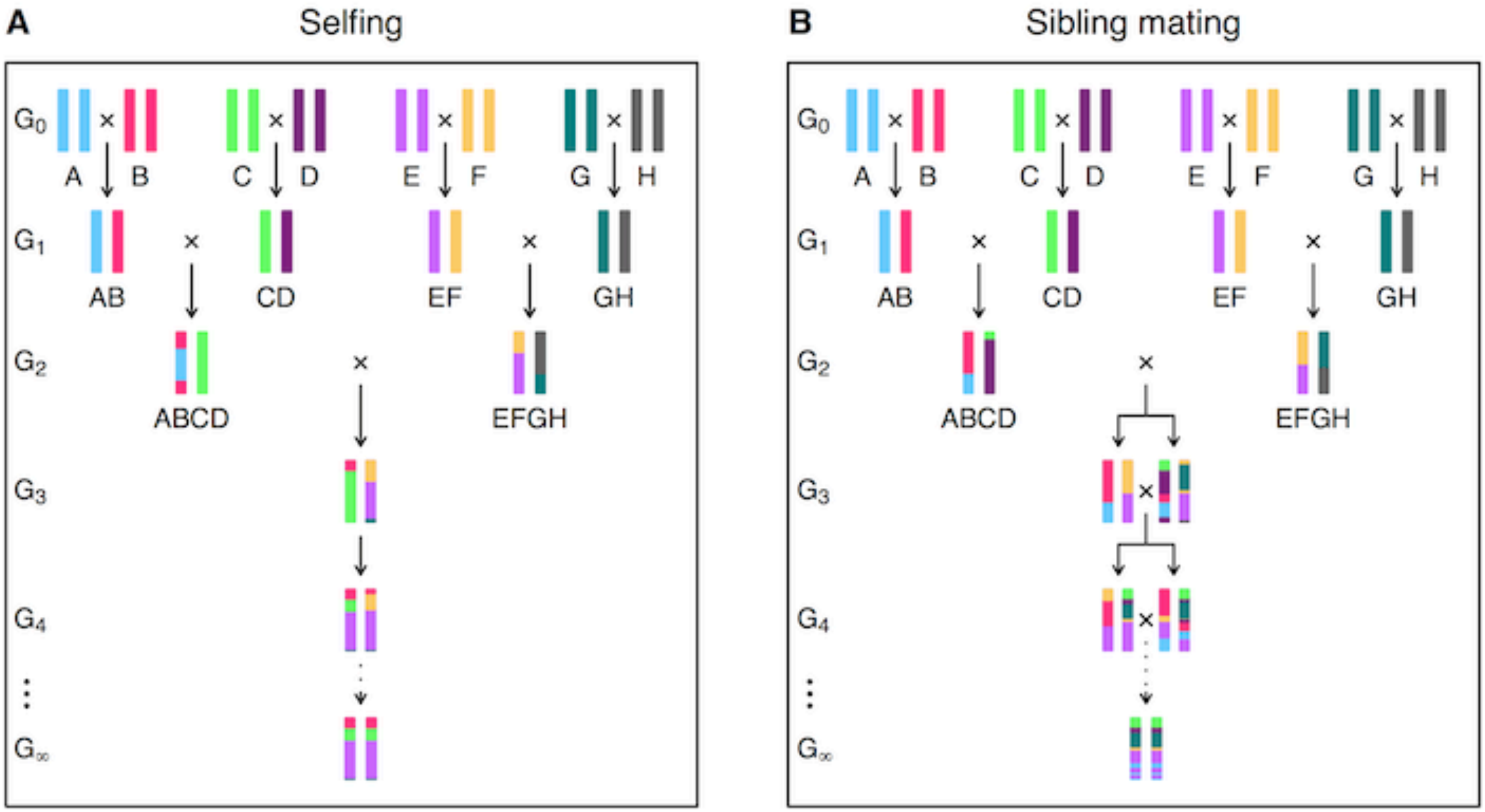
- 2 or more inbred founders
- single F1 self-pollinated
- generations of random mating
- generations of selfing
- aim for homozygosity at all loci

www.nature.com/nrg/journal/v9/n3/images/nrg2291-f4.jpg



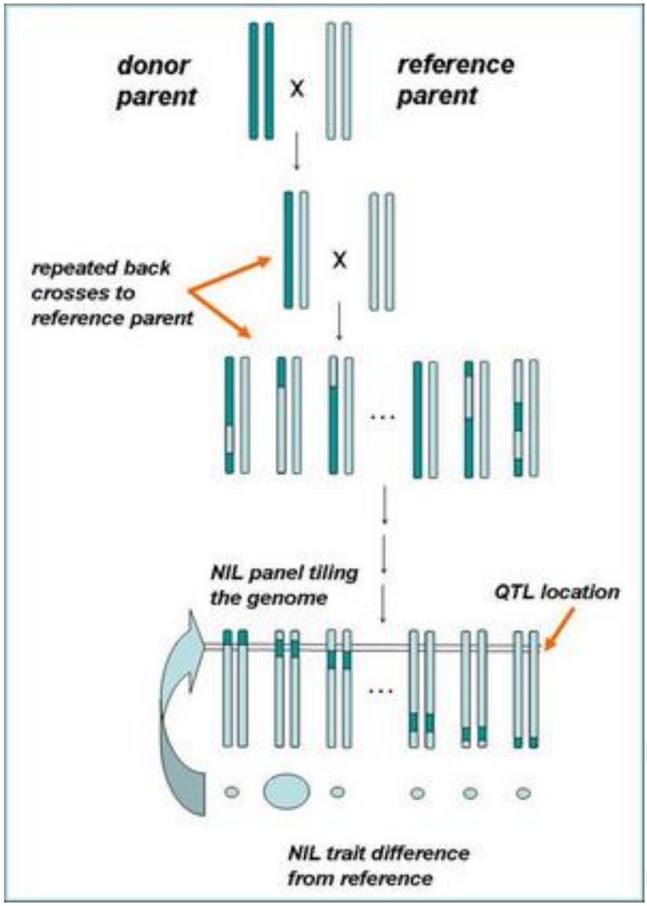
B73
×
Mo17
Single F1 plant was selfed
⊗ (self pollination)
Grow 200 plants, random mating
After a few more rounds of intermating, generate -200 RILs by repeated selfing
⊗ (self pollination)
Each RIL is examined for a particular trait, such as plant height. The genetic regions responsible for the trait can be mapped to a few cM
Nature Reviews | Genetics

Selfing vs sib mating

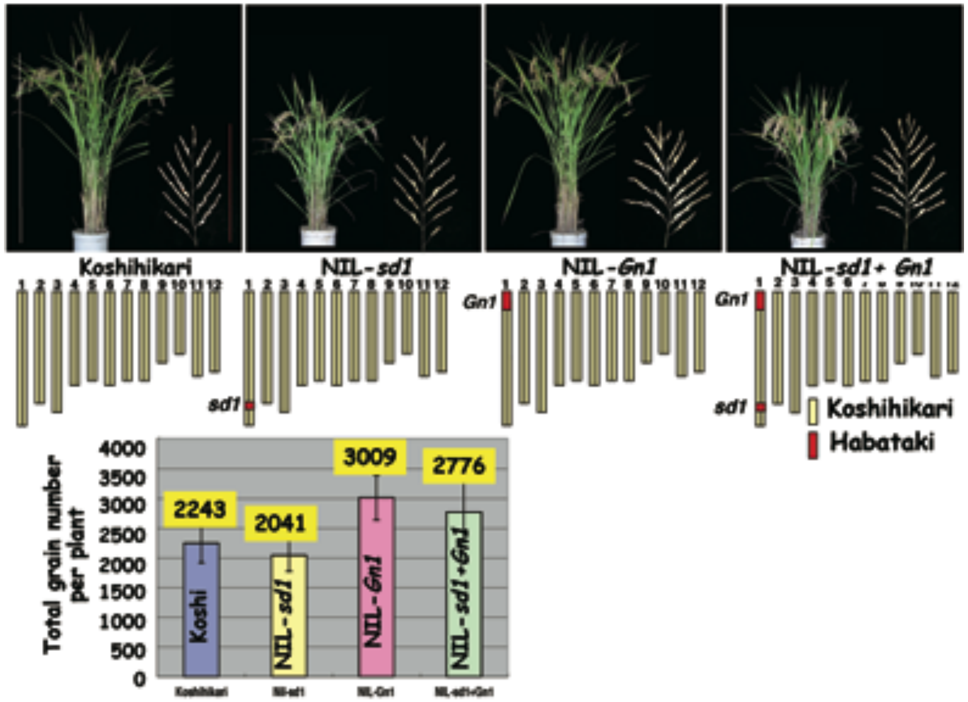


Broman (2005) Genetics

near isogenic lines (NIL)


donor parent
x
reference parent
repeated back crosses to reference parent
x
...
NIL panel tiling the genome
QTL location
...
NIL trait difference from reference

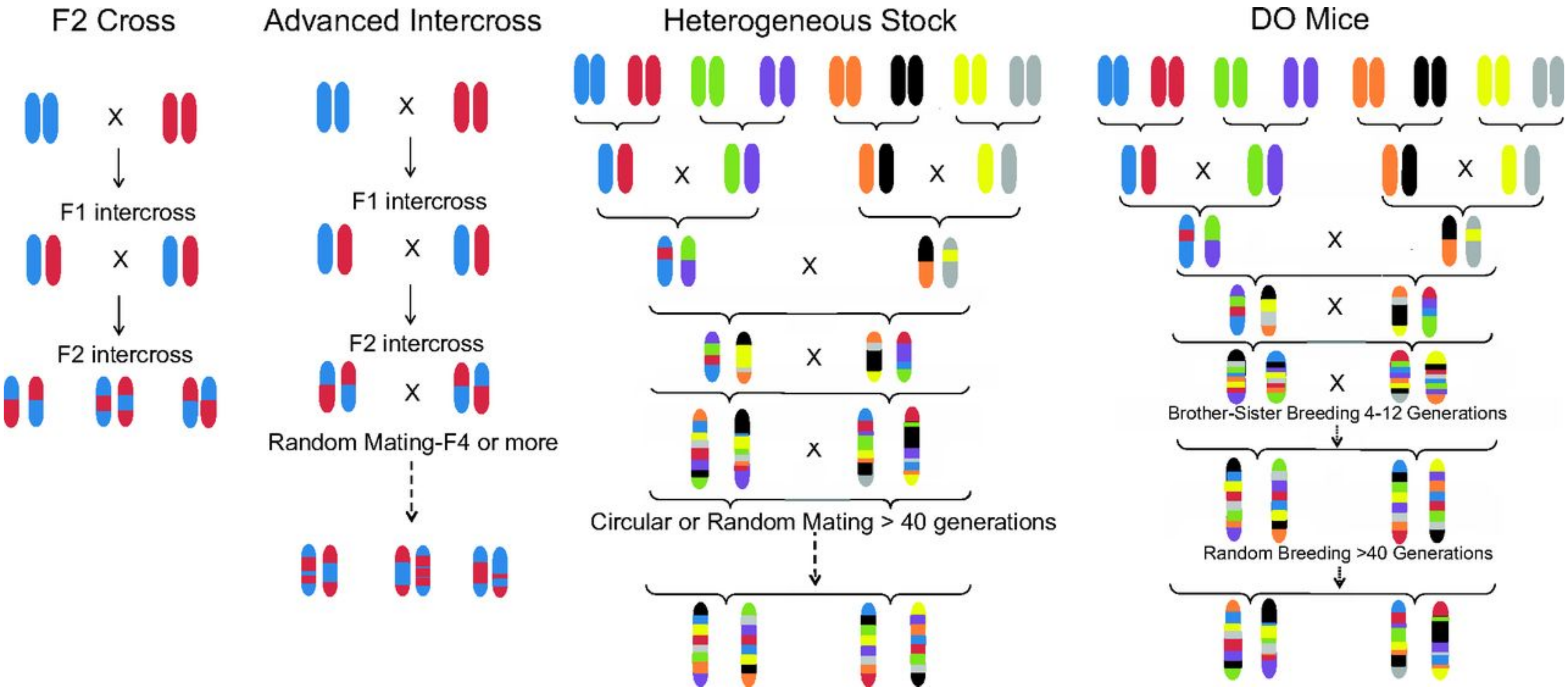
Rebecca Nelson



blog.generationcp.org/category/women-in-science-2/

meddic.jp/isogenic_line

Advanced Intercrosses

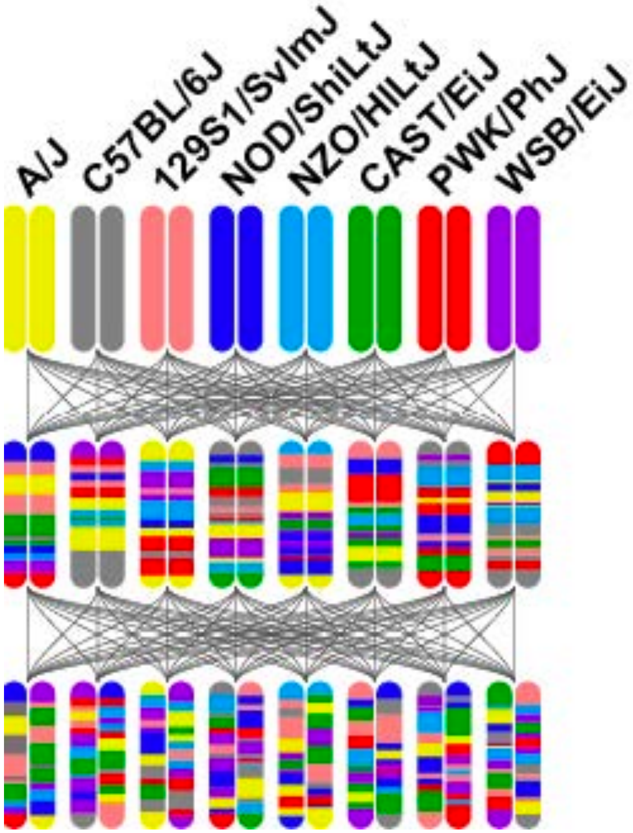

F2 Cross
Advanced Intercross
Heterogeneous Stock
DO Mice
x
F1 intercross
F2 intercross
Random Mating-F4 or more
Circular or Random Mating > 40 generations
Brother-Sister Breeding 4-12 Generations
Random Breeding >40 Generations

Leah Solberg Woods doi:10.1152/physiolgenomics.00127.2013

multi-parent populations

- more than 2 inbred parents (4,8,20)
- developed over generations
    - generations of cross-breeding
    - generations of selfing (or sibs)
- increased meiotic events
    - fine mapping to small region
    - SNP level in one generation

Laura Vanderploeg, Jackson Labs



natural populations?

- are genetic markers location on map?
    - marker analysis only?
    - local linkage disequilibrium
    - benefits of linkage analysis
- do rare alleles affect phenotype?
    - power depend on rare allele frequency
    - uneven inoformation across markers
- multi-parent populations capture useful diversity

dataset used in this talk

- Tom Osborn *Brassica napus* intercross (F2)
- Edgar Spalding *Arabidopsis thaliana* advanced intercrosses
- *Mus musculus* Diversity Outbred (DO)
    - Elissa Chesler & collaborators (Recla et al. 2014)
    - Alan Attie & collaborators (in progress)

Osborn *Brassica napus* intercross

- 104 doubled haploid (DH) lines
- 300 markers on 19 chromosomes
  - originally scattered linkage groups
- 9 phenotypes (flowering time & seedling survival)

Ferreira, Satagopan, Yandell, Williams, Osborn (1995) TAG

Satagopan, Yandell, Newton, Osborn (1996) Genetics

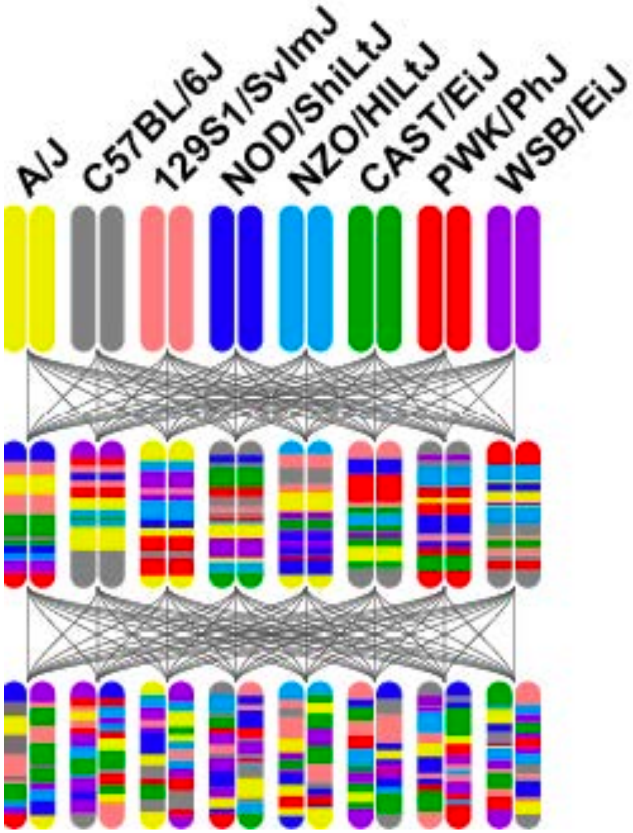
Moore/Spalding *A. thaliana* NIL & RILs

- *Arabidopsis thaliana* Ler x Cvi population
    - 92 near-isogenic lines (NIL); 2525 seedlings
    - 162 RILs; 2132 (RIL1) or 2325 (RIL2) seedlings
- genotypes: 102 (NIL) or 234 (RILs) markers on 5 chr
- phenotypes: 241 root tip angles, every 2 min
- automated image acquisition & analysis
    - images: 7000 lines x 241 time points
    - genome scans across all time points
- botany / computer science / biostatistics collaboration

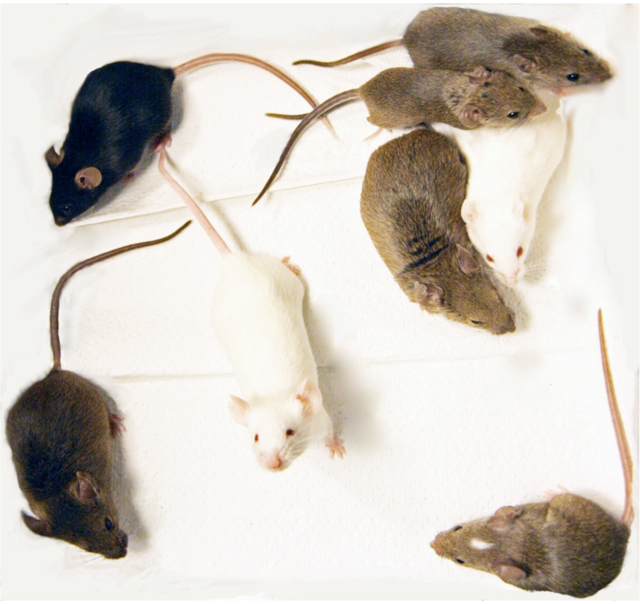
Moore, Johnson, Kwak, Livny, Broman, Spalding (2013)

Diversity Outbred example

- 283 mice (generations 4 & 5)
- 320 (of 7851) SNP markers
- phenotype = `OF_immobile_pct` (of 1000s)
- Data: https://github.com/rqtl/qtl2data/
- Recla, Robledo, Gatti, Bult, Churchill, Chesler (2014)



Attie/Jax DO population



- 8 CC founder strains (generation 19-22)
- 500 mice in 5 waves
- multiple traits measured
    - 150K SNP GIGA-MUGA chip imputed to 40M SNPs
    - 100s clinical traits (insulin secretion)
    - 30K RNA-Seq expression traits
    - 2K proteomic, 200 metabolomic, 200 lipidomic
    - microbiome: 2K of 16s; 1M of sequencing