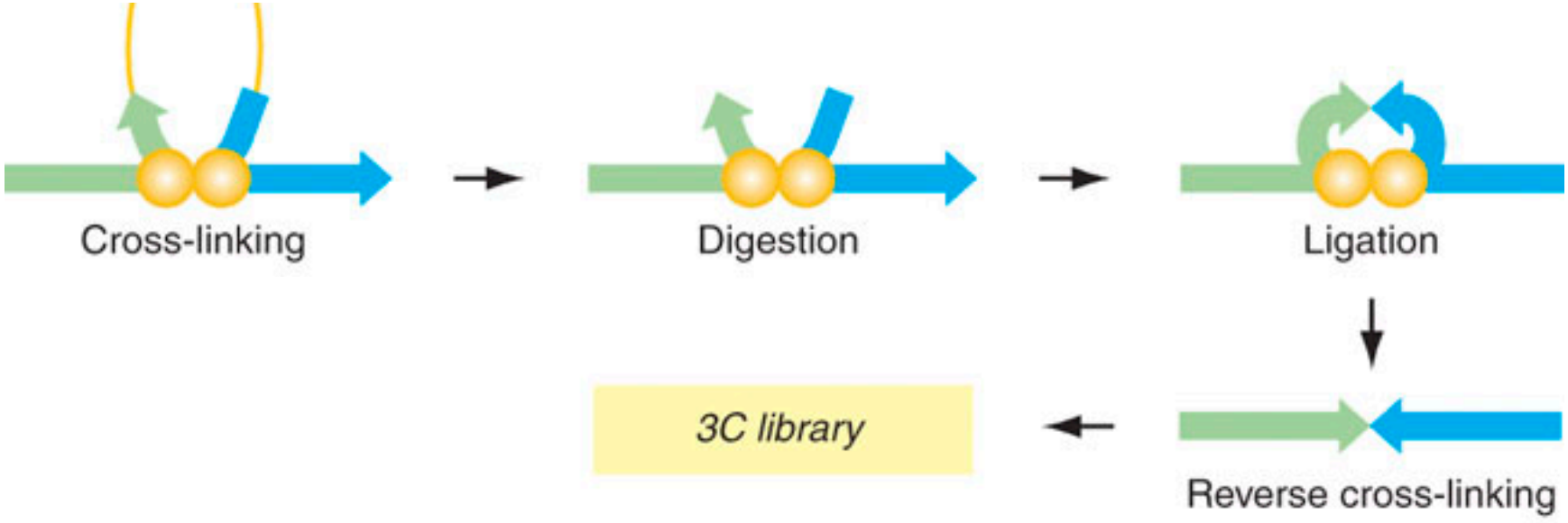
Quantitative Trait Loci (QTLs)
Lecture 19

David K. Gifford

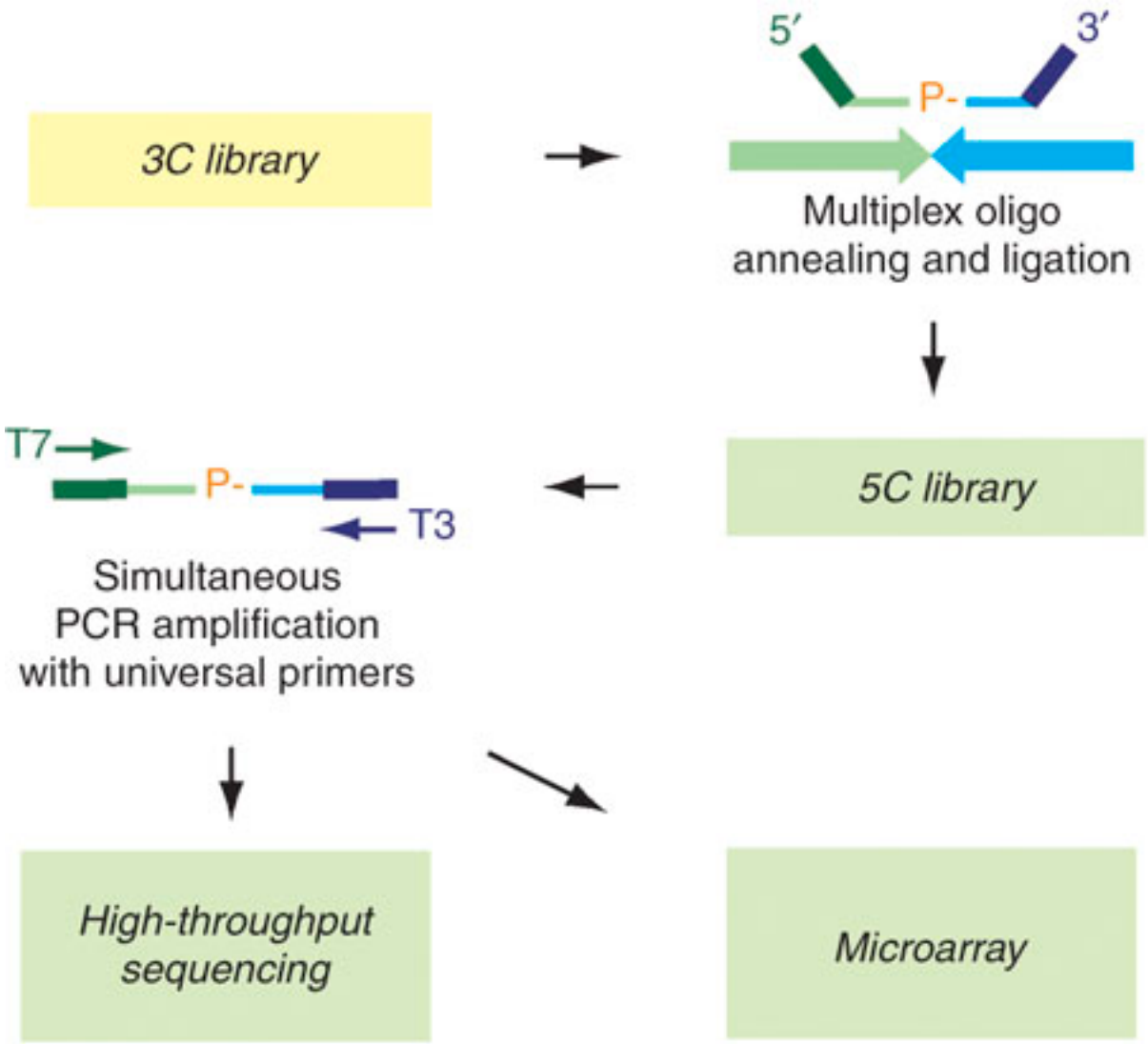
Massachusetts Institute of Technology

5C maps interactions between defined primers

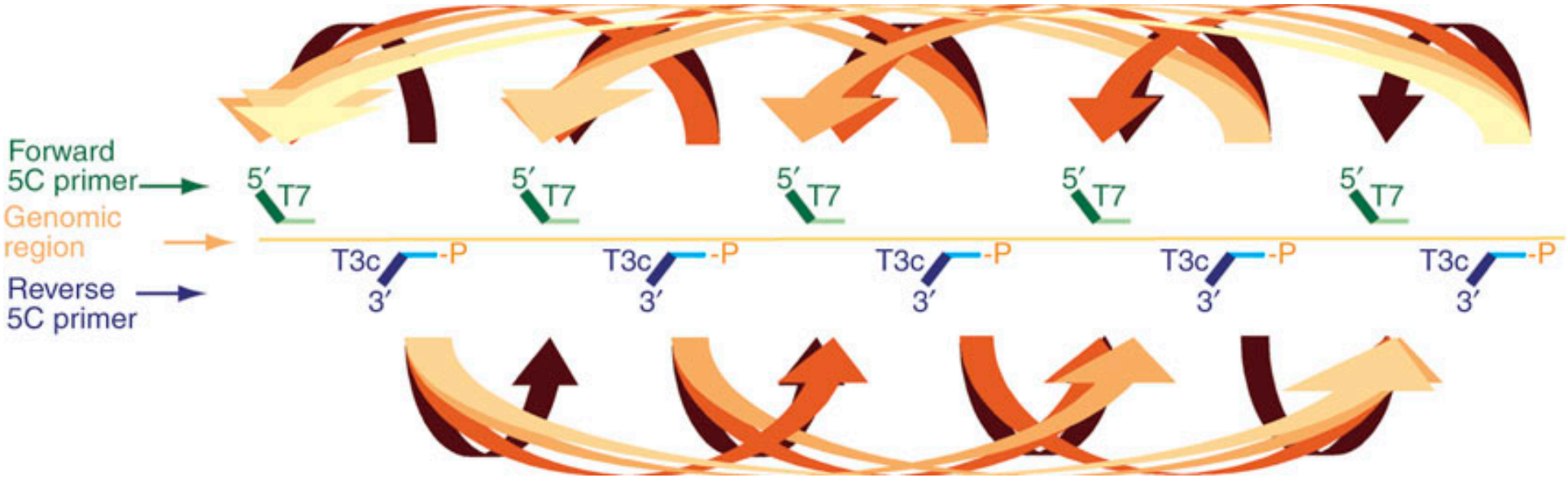


Source: Dostie, Josée, and Job Dekker. "Mapping Networks of Physical Interactions Between Genomic Elements using 5C Technology." *Nature Protocols* 2, no. 4 (2007): 988-1002.

5C maps interactions between defined primers



5C maps interactions between defined primers



Courtesy of Macmillan Publishers Limited. Used with permission.
Source: Dostie, Josée, and Job Dekker. "Mapping Networks of Physical Interactions Between Genomic Elements using 5C Technology." *Nature Protocols* 2, no. 4 (2007): 988-1002.

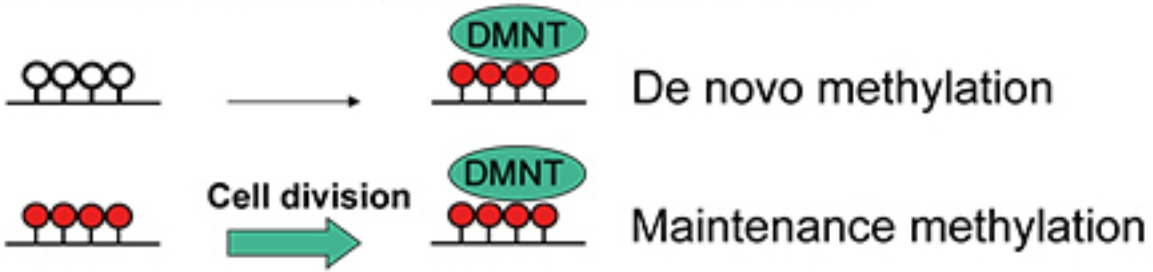
DNA methylation

Addition of a methyl group to a cytosine within C-G di-nucleotides which are frequently located in the regulatory regions of genes.

A mechanism for gene silencing:

- ⇨ preventing binding of regulatory factors
- ⇨ affecting chromatin status

CpG methylation is a lasting form of epigentic modification



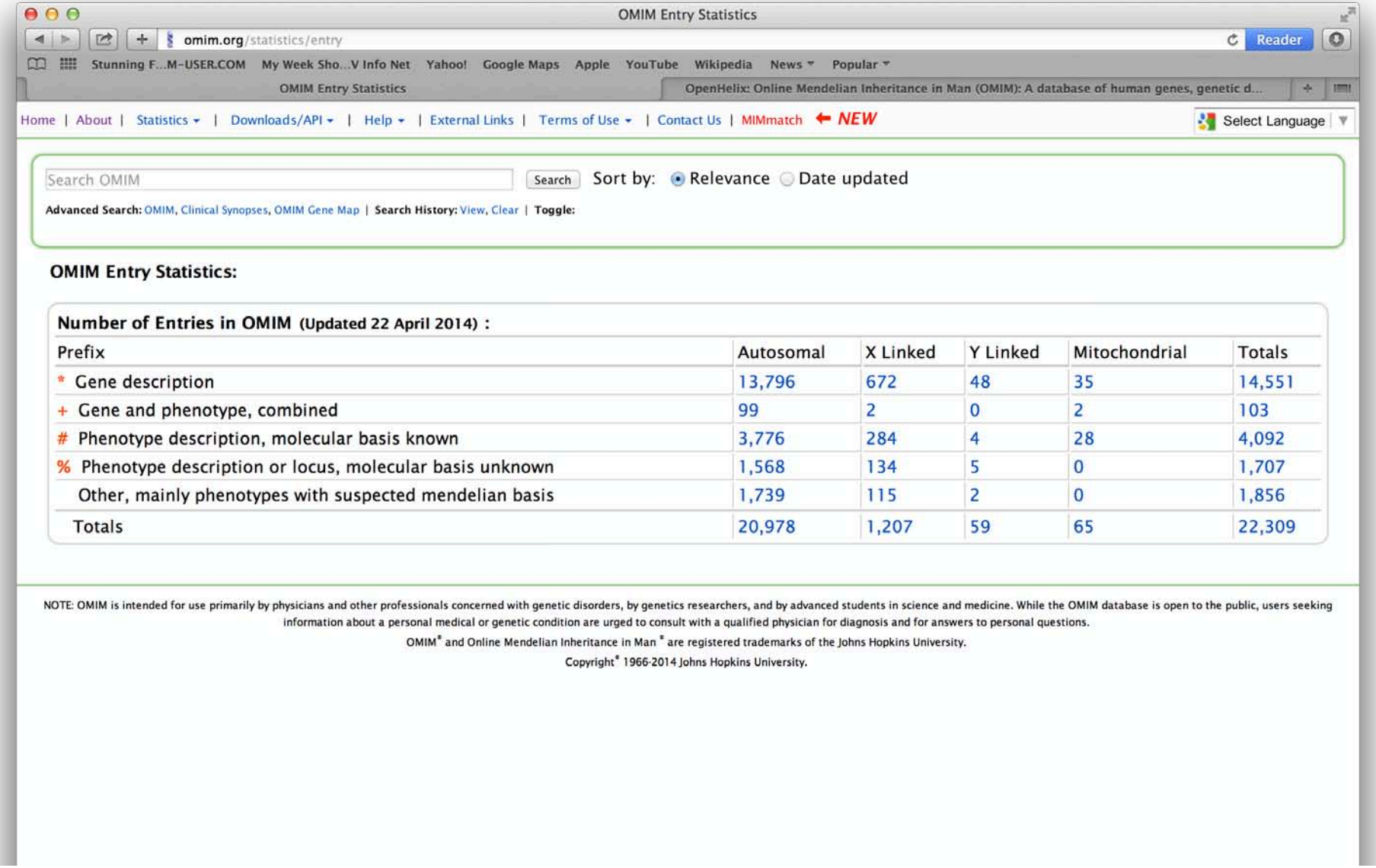
Today's Narrative Arc

1. Usually, you are more like your relatives than random people on the planet.
2. The heritability of a trait is the fraction of phenotypic variance that can be explained by genotype
3. Computational models that predict phenotype from genotype are key for understanding disease related genomic variants and the most effective therapy for a disease (pharmacogenomics)
4. We will computationally predict quantitative phenotypes by adding the contribution of individual loci (QTLs)
5. Typically our models can only predict a small fraction of phenotypic variance – the so called "missing heritability" problem

Today's Computational Approaches

1. Linear models of phenotype that use stepwise regression and forward feature selection
2. Test statistics for discovering significant QTLs
3. Measurement of narrow sense heritability (h^2), broad sense heritability (H^2), and environmental variance

OMIM - authoritative compendium of human genes and genetic phenotypes related to Mendelian Inheritance



OMIM Entry Statistics

omim.org/statistics/entry

Home | About | Statistics | Downloads/API | Help | External Links | Terms of Use | Contact Us | MIMmatch NEW

Search OMIM Search Sort by: Relevance Date updated

Advanced Search: OMIM, Clinical Synopses, OMIM Gene Map | Search History: View, Clear | Toggle:

OMIM Entry Statistics:

Number of Entries in OMIM (Updated 22 April 2014) :

Prefix	Autosomal	X Linked	Y Linked	Mitochondrial	Totals
* Gene description	13,796	672	48	35	14,551
+ Gene and phenotype, combined	99	2	0	2	103
# Phenotype description, molecular basis known	3,776	284	4	28	4,092
% Phenotype description or locus, molecular basis unknown	1,568	134	5	0	1,707
Other, mainly phenotypes with suspected mendelian basis	1,739	115	2	0	1,856
Totals	20,978	1,207	59	65	22,309

NOTE: OMIM is intended for use primarily by physicians and other professionals concerned with genetic disorders, by genetics researchers, and by advanced students in science and medicine. While the OMIM database is open to the public, users seeking information about a personal medical or genetic condition are urged to consult with a qualified physician for diagnosis and for answers to personal questions.

Courtesy of Johns Hopkins University. Used with permission.

Statistics review

$$\mu_x = \frac{1}{N}\sum_{i=1}^{N} x_i$$

$$\sigma_x^2 = \frac{1}{N}\sum_{i=1}^{N}\left(x_i - \mu_x\right)^2 = E\left[\left(X - \mu_x\right)^2\right]$$

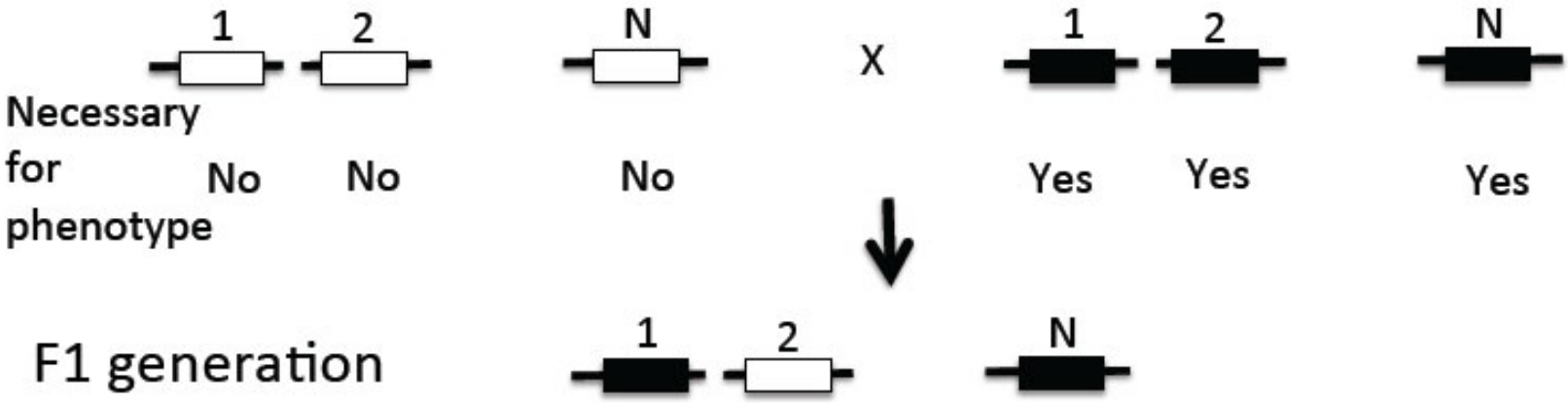
$$\sigma_{xy}^2 = E\left[\left(X - \mu_x\right)\left(Y - \mu_y\right)\right]$$

Covariance
=0 when X and Y are independent

Genotype to Phenotype

- Genotype
  - Complete genome sequence (or an approximation)
  - Can be defined by markers at specific genomic sites that describe differences with a defined reference genome
- A phenotype is defined by one or more traits
- Non-quantitative trait (dead/alive, etc.)
- Quantitative Trait
  - Fitness (growth rate, lifespan, etc.)
  - Morphology (height, etc.)
  - Gene expression
- Quantitative Trait Loci - Marker that is associated with a quantitative trait
  - eQTL – marker associated with gene expression

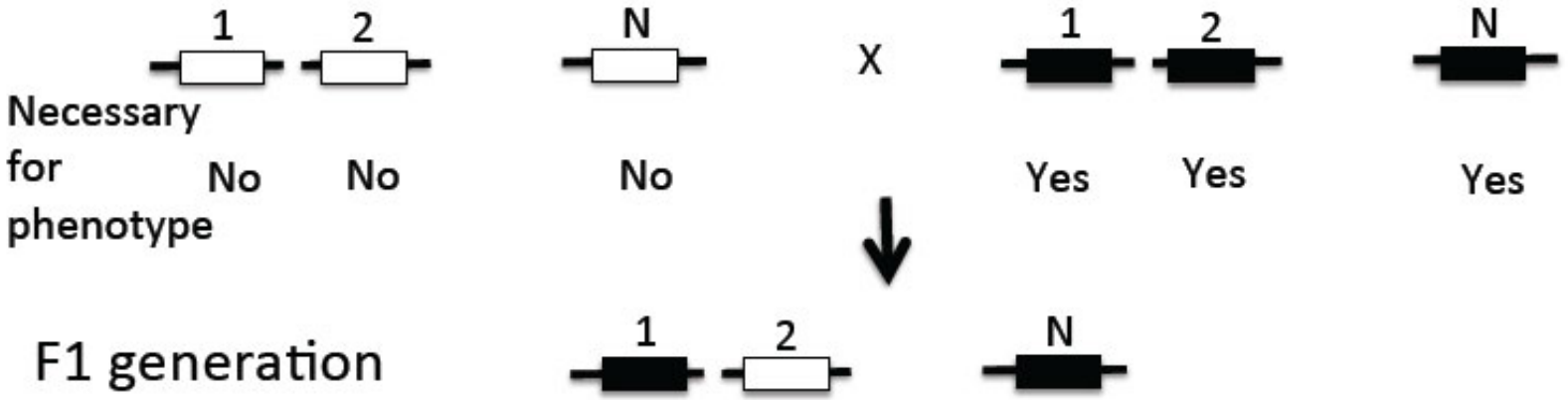
Binary haploid genetic model



Example Phenotype

Alive/Dead in a specific environment

Binary haploid genetic model

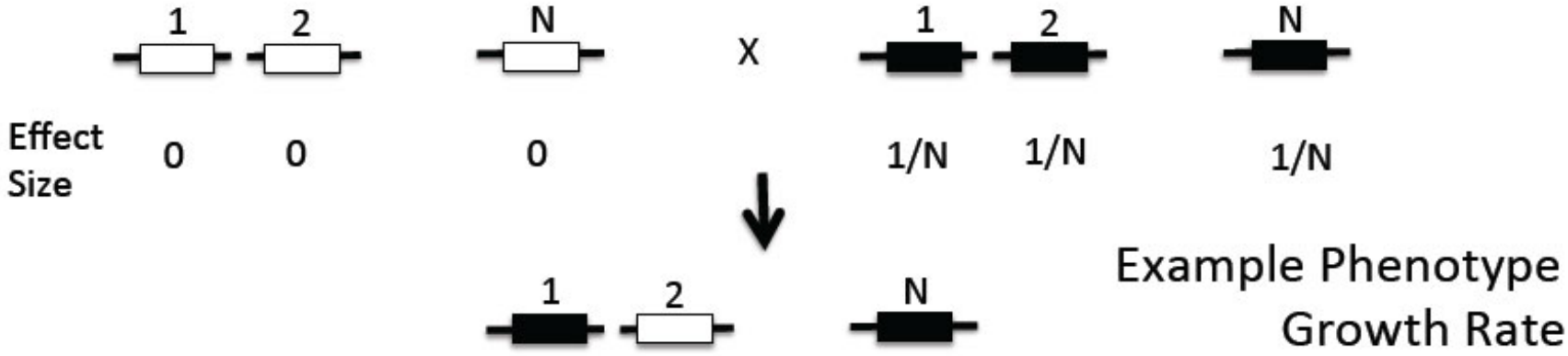


Example Phenotype

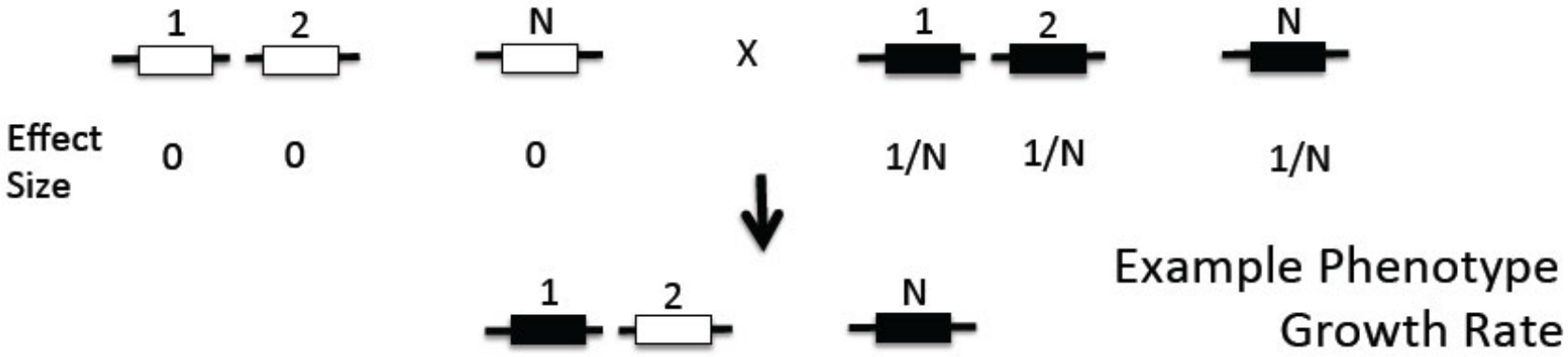
Alive/Dead in a specific environment

N is estimated by $\log_2$ (# F1s tested / # F1s with phenotype)

Quantitative haploid genetic model



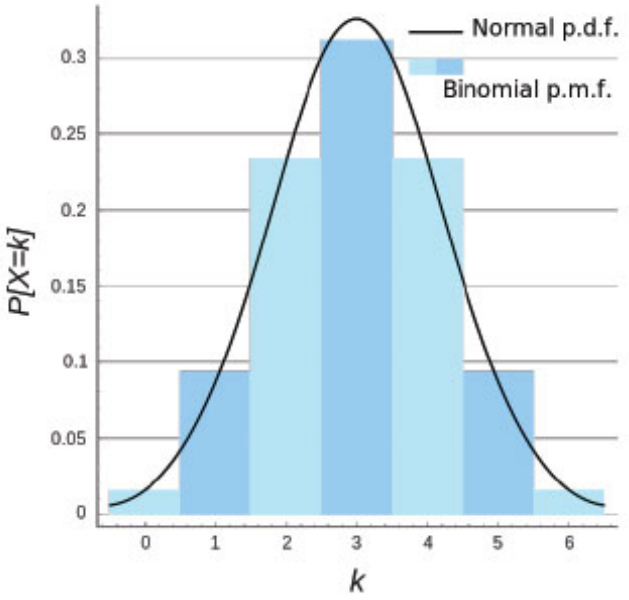
Quantitative haploid genetic model



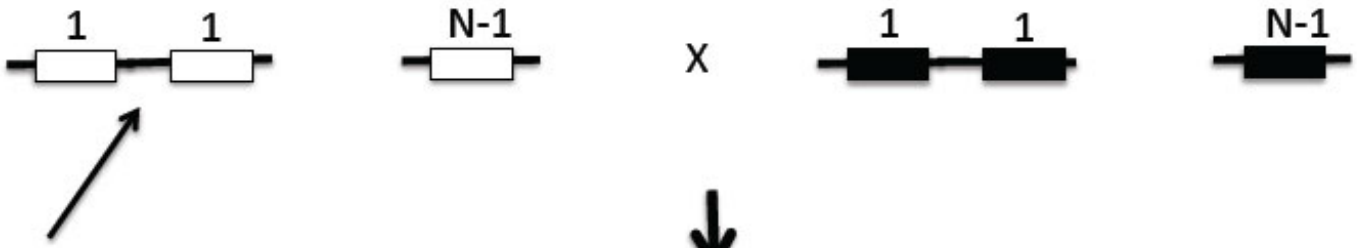
$$p(x,N) = \binom{N}{x}(1-.5)^{N-x}.5^{x}$$

$$E[x] = .5$$

$$\sigma_x^2 = .25 / N$$



Genetic linkage causes marker correlation



Proximal genomic locations makes crossing over unlikely during meiosis

Phenotype is a function of genotype plus an environmental component

- i – individual in [1 .. N]
- g_i – genotype of individual i
- p_i – quantitative phenotype of individual i (single trait)
- e_i – environmental contribution to p_i

Phenotype is a function of genotype plus an environmental component

- i – individual in [1 .. N]
- g_i – genotype of individual i
- p_i – quantitative phenotype of individual i (single trait)
- e_i – environmental contribution to p_i

$$p_i = f(g_i) + e_i \qquad \sigma_p^2 = \frac{1}{N}\sum_{i=1}^{N}\left(p_i - \mu_p\right)^2$$

$$\sigma_p^2 = \sigma_g^2 + \sigma_e^2 + 2\sigma_{ge}^2 \qquad E[e_i] = 0 \qquad E[e^2] = \sigma_e^2$$

g and e assumed or made independent yields $\sigma_p^2 = \sigma_g^2 + \sigma_e^2$

Why two heritabilities?

- Broad-sense
  - Describes the upper bound for phenotypic prediction by an optimal arbitrary model
  - Reveals complexity of molecular mechanism
- Narrow-sense
  - Describes the upper bound for phenotypic prediction by a linear model
  - Describes relative resemblance and utility of family disease history
  - Efficienct genetic mapping studies

Key caveats

- Heritability is a property of population (segregating allele frequencies) and environment (noise component)
- "Heritability" in practice may refer to either broad- or narrow-sense (or an implicit assumption that they are the same)
- Estimation is difficult (matching environments and avoiding confounding)

H^2 - Broad Sense heritability

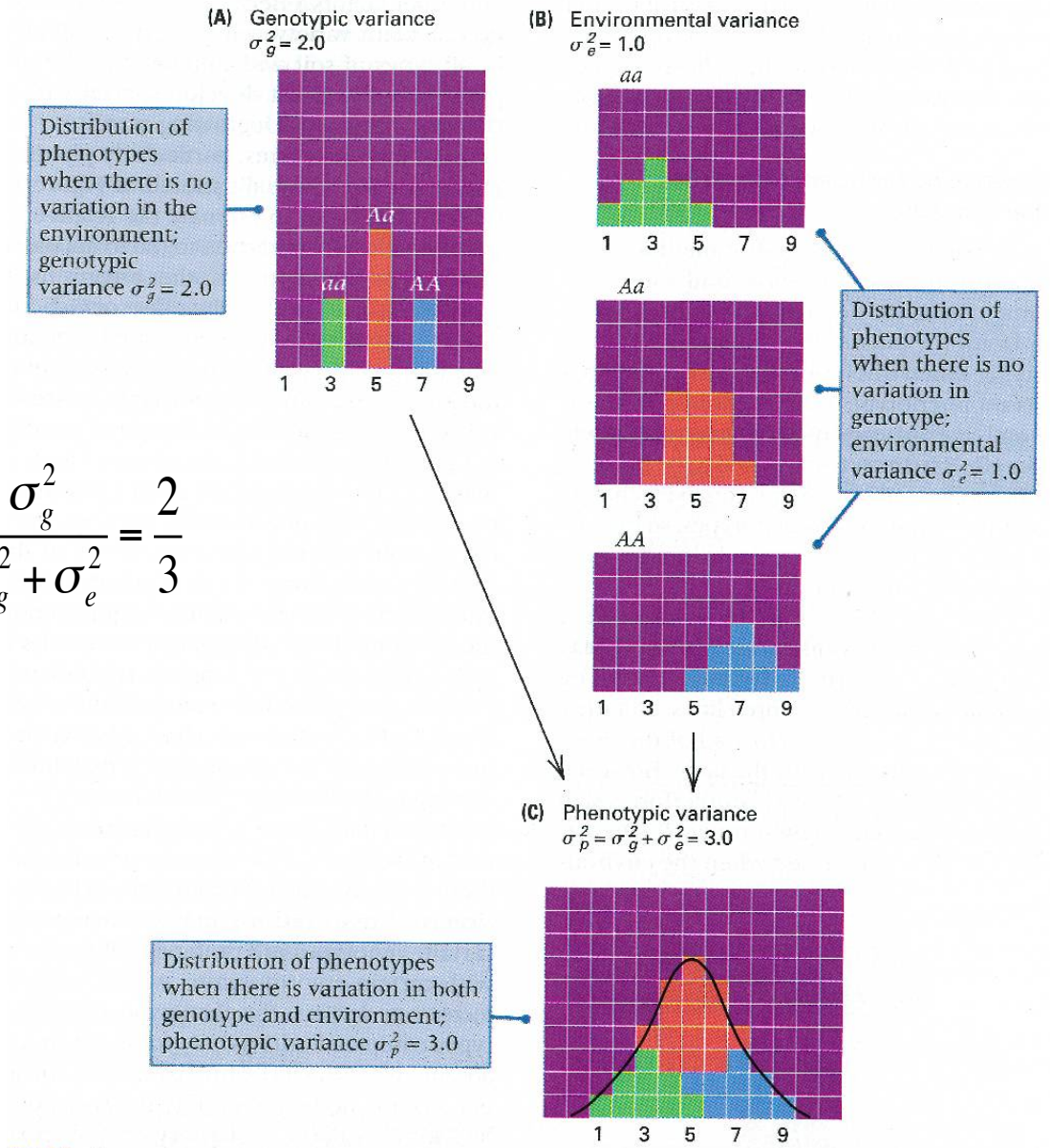
- Fraction of phenotypic variance explained by genetic component

$$H^2 = \frac{\sigma_g^2}{\sigma_p^2} = \frac{\sigma_p^2 - \sigma_e^2}{\sigma_p^2}$$

- Can estimate σ_e^2 from identical twins or clones.

Broad heritability of a trait is fraction of phenotypic variance explained by genetic causes

$$H^2 = \frac{\sigma_g^2}{\sigma_p^2} = \frac{\sigma_g^2}{\sigma_g^2 + \sigma_e^2} = \frac{2}{3}$$



Source: Hartl, Daniel L. Essential Genetics: A Genomics Perspective. Jones & Bartlett Publishers, P. 506, 2011.

Additive model of phenotype

g_{ij} is marker j for individual i with values {0,1}
Quantitative trait loci (QTLs) are discovered for each trait

$$f_a(g_i) = \sum_{j \in QTL} \beta_j g_{ij} + \beta_0$$

$$E\left[f_a(g_i)\right] = \frac{f_a(p_1)}{2} + \frac{f_a(p_2)}{2}$$

Children tend to midpoint of parents for additive traits as they are expected to get an equal number of loci from each parent

Historical heritability example

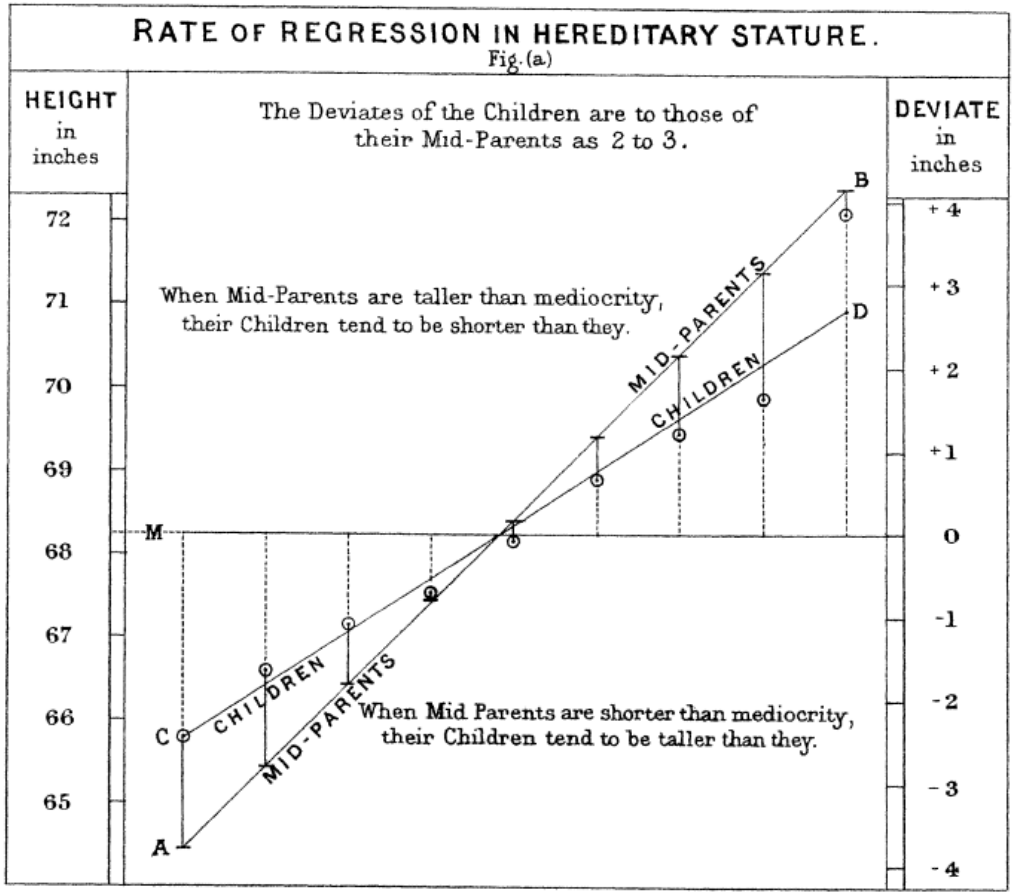


Figure is in the public domain.

Galton, "Regression towards mediocrity in hereditary stature" (1886)

h^2 - Narrow Sense heritability

- Fraction of phenotypic variance explained by an additive model of markers
- $f_a(g_i)$ is additive model of genotypic components in g_i
- Difference between heritability explained by additive model and general model is one source of “missing heritability” in current studies

h^2 - Narrow Sense heritability

- Fraction of phenotypic variance explained by an additive model of markers
- $f_a(g_i)$ is additive model of genotypic components in g_i
- Difference between heritability explained by additive model and general model is one source of "missing heritability" in current studies

$$p_i = f_a(g_i) + e_i \qquad \sigma_a^2 = \sigma_p^2 - \frac{1}{N}\sum_{i=1}^{N}\left(p_i - f_a(g_i)\right)^2$$

$$h^2 = \frac{\sigma_a^2}{\sigma_p^2}$$

Example trait heritabilities – h^2

Morphological Traits

- Human height ~ .8
- Cattle Yearling Weight ~ .35

Fitness Traits

- Drosophila life history ~ .2
- Wild animal life history ~ .3

h^2 from Visscher et al. 2008

Example trait heritabilities

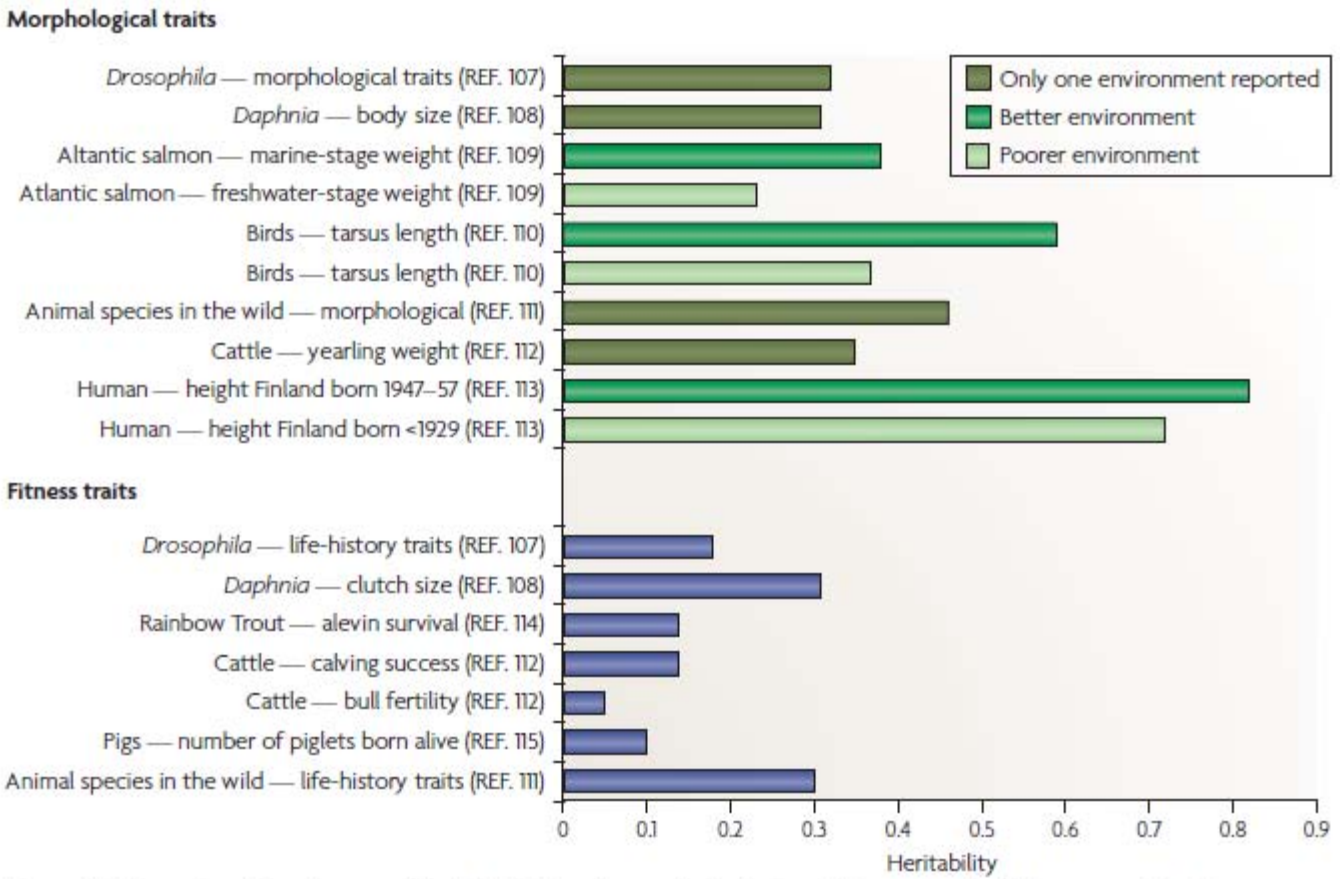


Figure 1 | **Examples of estimates of heritabilities of morphological and fitness traits.** Where possible, the estimates of heritability were taken from Reviews, and are the mean across a number of studies. The examples show that, on average, heritability estimates are larger for morphological traits than for fitness-related traits, and that heritability tends to be larger in better environments when compared with poorer environments.

h^2 from Visscher et al. 2008

Today's Narrative Arc

1. Usually, you are more like your relatives than random people on the planet.
2. The heritability of a trait is the fraction of phenotypic variance that can be explained by genotype
3. **Computational models that predict phenotype from genotype are key for understanding disease related genomic variants and the most effective therapy for a disease (pharmacogenomics)**
4. We will computationally predict quantitative phenotypes by adding the contribution of individual loci (QTLs)
5. Typically our models can only predict a small fraction of phenotypic variance – the so called "missing heritability" problem

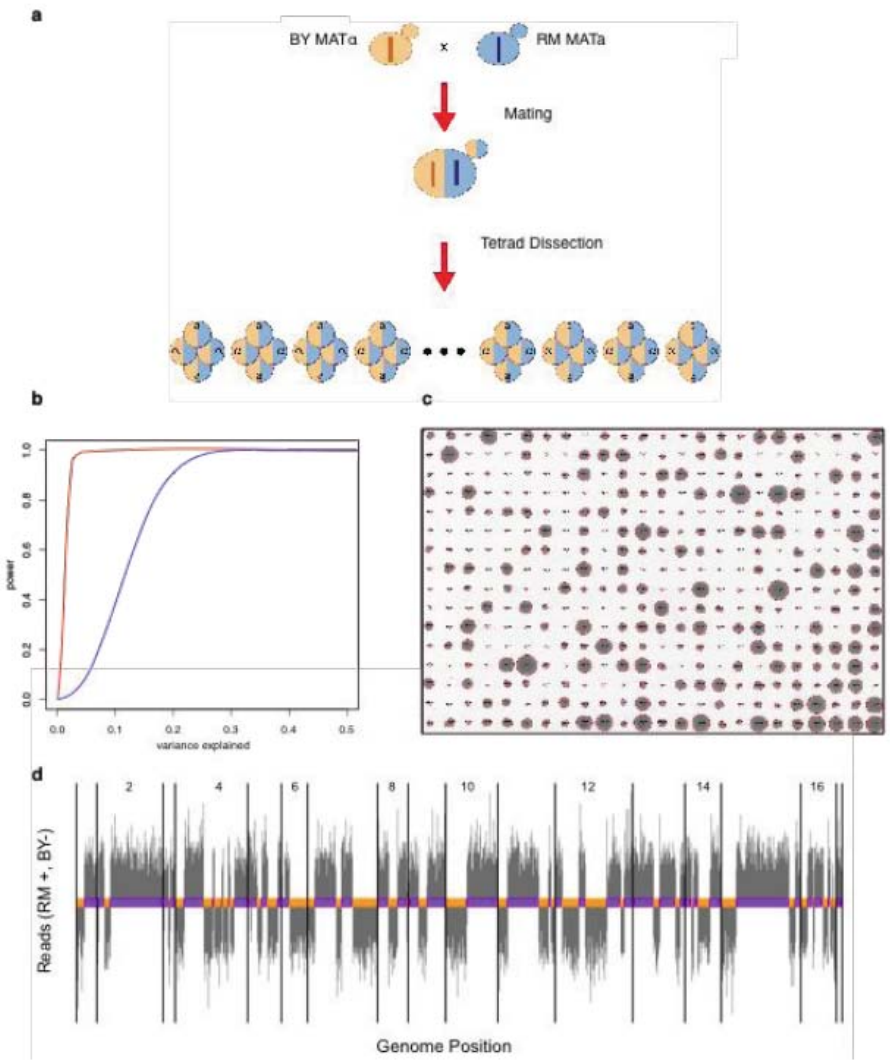
Can we predict phenotype in a haploid yeast system?

Finding the sources of missing heritability in a yeast cross

Joshua S. Bloom[1,2], Ian M. Ehrenreich[1,3], Wesley T. Loo[1,2], Thúy-Lan Võ Lite[1,2] & Leonid Kruglyak[1,4,5]

NATURE | VOL 494 | 14 FEBRUARY 2013

Study heritability of 46 traits in ~1000 segregants



Courtesy of Macmillan Publishers Limited. Used with permission.
Source: Bloom, Joshua S., Ian M. Ehrenreich, et al. "Finding the Sources of Missing Heritability in a Yeast Cross." *Nature* 494, no. 7436 (2013): 234-7.

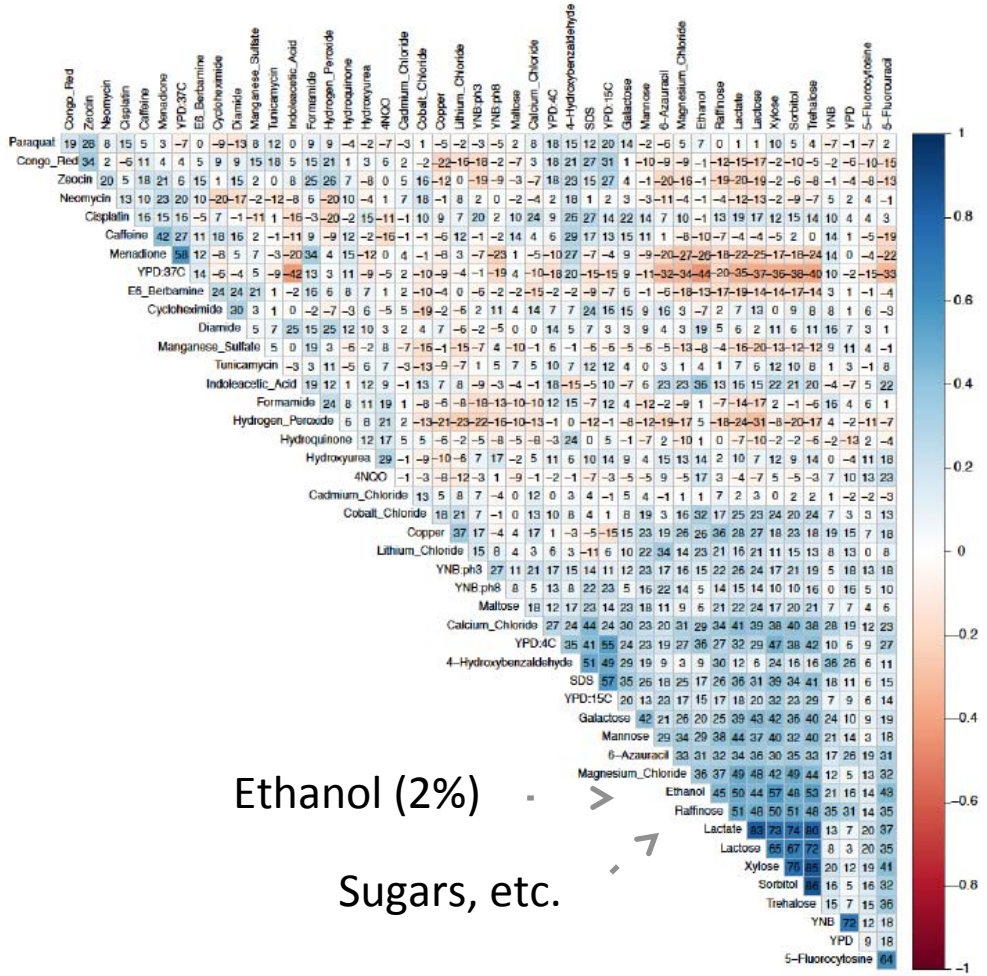
Key advance: large panel (~1000 segregants), many phenotypes (46)

BY and RM parents

Figure S1. The design of the segregant panel is shown in (A). (B) Curves illustrating statistical power are shown for mapping populations of 100 (blue) and 1000 (red) segregants at a genome-wide significance threshold. (C) An image of endpoint colony growth is shown for 384 segregants, with the outlines of colonies, as detected by our image processing software, indicated in red. (D) Counts of sequencing reads at SNP sites are plotted (Y-axis) against genome position (X-axis) for a representative segregant; the orange (BY) and purple (RM) bars indicate parental haplotype calls, and the vertical black bars delineate chromosomes.

Bloom et al. 2013

Certain phenotypes are related



Bloom et al. 2013

Figure S2. Spearman correlation coefficients for all pairs of traits are shown. Numbers in table cells indicate (100 * correlation coefficient).

Source: Bloom, Joshua S., Ian M. Ehrenreich, et al. "Finding the Sources of Missing Heritability in a Yeast Cross." *Nature* 494, no. 7436 (2013): 234-7.

Today's Narrative Arc

1. Usually, you are more like your relatives than random people on the planet.
2. The heritability of a trait is the fraction of phenotypic variance that can be explained by genotype
3. Computational models that predict phenotype from genotype are key for understanding disease related genomic variants and the most effective therapy for a disease (pharmacogenomics)
4. **We will computationally predict quantitative phenotypes by adding the contribution of individual loci (QTLs)**
5. Typically our models can only predict a small fraction of phenotypic variance – the so called "missing heritability" problem

LOD scores to discover QTLs

$$LOD = \log_{10} \prod_{i=1}^{N} \frac{P\left(p_i \mid g_{ij}, \mu_0, \mu_1, \sigma\right)}{P\left(p_i \mid \mu, \sigma\right)}$$

- Use trait means conditioned on marker j in individual vs. unconditioned mean for trait to test if marker j is a QTL
- Permute genotypes 1000 times and each time compute LOD scores to estimate null LOD distribution
- Determine null LOD score that describes FDR = 0.05
- Use this threshold on unpermuted LOD scores to find QTLs for each gene
- Fit linear model to discovered QTLs
- Repeat finding QTLs predicting residuals from existing model (3 times)

1005 segregants detect more QTLs than 100 segregants

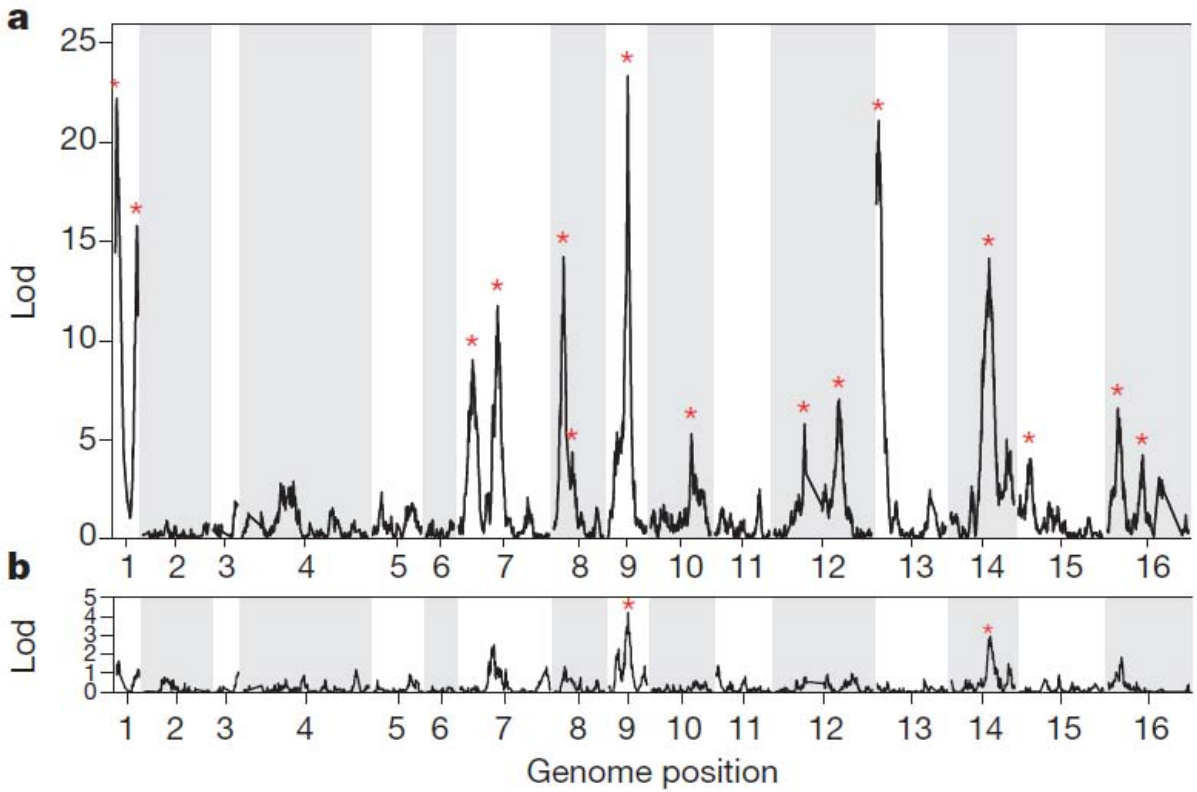


Figure 3 | QTL detection for a complex trait. Lod score is plotted against the genetic map. Red asterisks indicate statistically significant QTL. **a**, Lod score plot with 1,005 segregants for growth in E6 berbamine. **b**, Lod score plot with 100 segregants for growth in E6 berbamine. The 15 significant QTL in **a** explain 78% of the narrow-sense heritability, compared with 21% for the 2 significant QTL in **b**. Alternating shaded bands denote chromosome boundaries.

Bloom et al. 2013

Phenotype prediction works well with identified QTLs

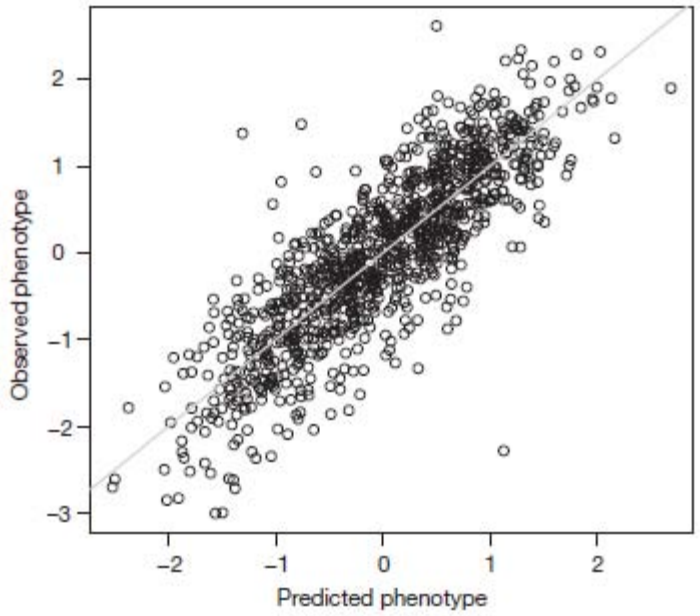
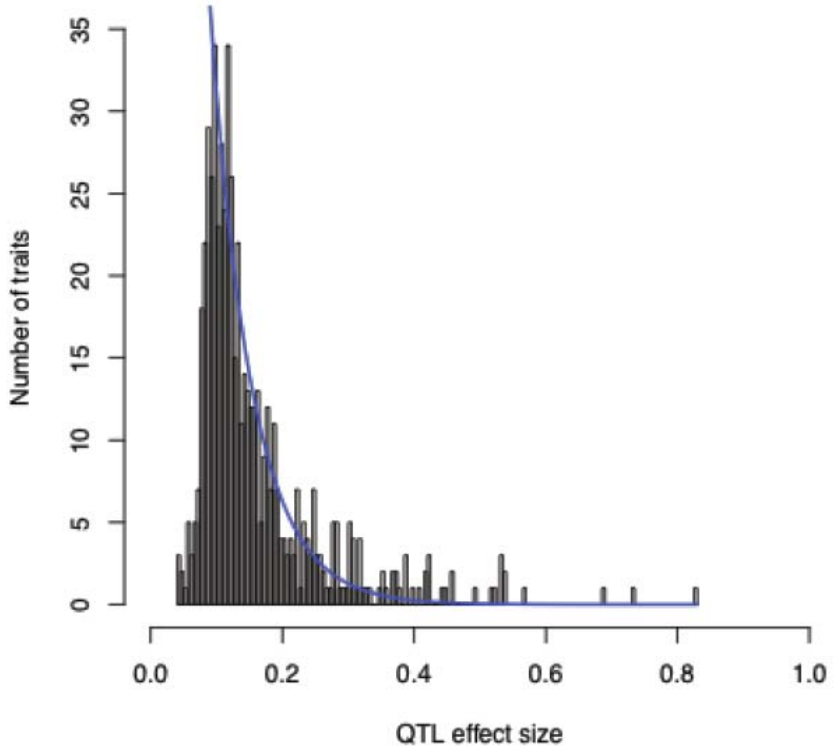


Figure 4 | Prediction of segregant trait values from QTL phenotypes. The observed phenotypic values for growth in lithium chloride are plotted against the predicted phenotypic values based on a cross-validated additive model of 22 QTL. The additive QTL model explains 88% of the narrow-sense heritability. The diagonal line represents (observed phenotype) = (predicted phenotype) and is shown as a visual guide.

Bloom et al. 2013

Most identified QTLs have small effects

5-29 QTLs per trait (median of 12), reported at 5% FDR

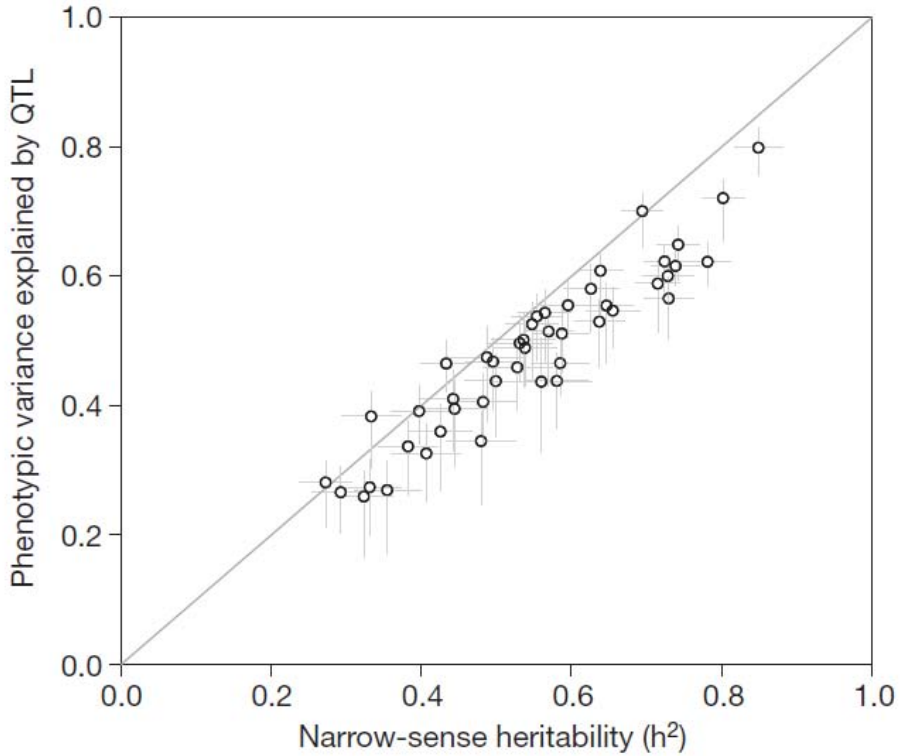


Bloom et al. 2013

Figure S3. A histogram of QTL effect sizes across all traits is plotted, showing that most detected QTL have small effects. Effect size here is the absolute value of the standardized difference in allelic means for each QTL. The blue line indicates a fit of a truncated exponential distribution of effect sizes.

Identified QTLs explain most additive heritability

QTLs explain 72-100% of narrow-sense heritability

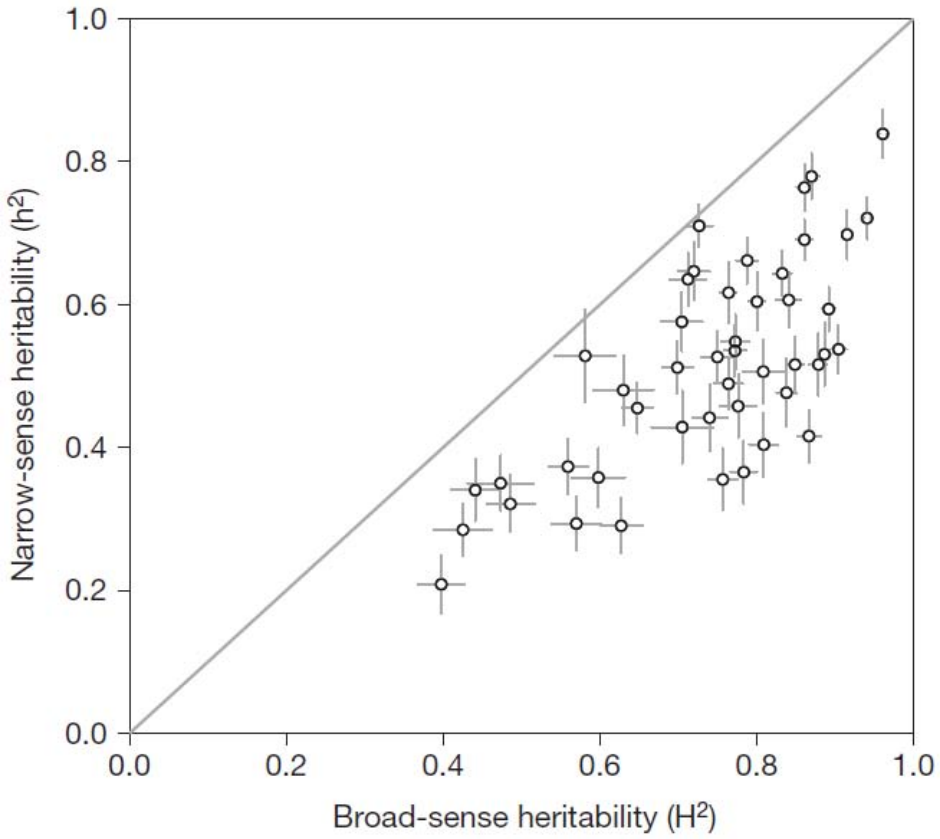


Bloom et al. 2013

Figure 2 | Most additive heritability is explained by detected QTL. **a**, The total variance explained by detected QTL for each trait is plotted against the narrow-sense heritability (h^2). Error bars show ± s.e. The diagonal line represents (variance explained by detected QTL) = h^2 and is shown as a visual guide.

"Missing Heritability" exists with our linear model

Vertical gap represents non-additive genetic contributions



Bloom et al. 2013

Figure 1 | Heritability for 46 yeast traits. The narrow-sense heritability (h^2) for each trait is plotted against the broad-sense heritability (H^2). Error bars show ± s.e. in heritability estimates. The diagonal line represents $h^2 = H^2$ and is shown as a visual guide.

Today's Narrative Arc

1. Usually, you are more like your relatives than random people on the planet.
2. The heritability of a trait is the fraction of phenotypic variance that can be explained by genotype
3. Computational models that predict phenotype from genotype are key for understanding disease related genomic variants and the most effective therapy for a disease (pharmacogenomics)
4. We will computationally predict quantitative phenotypes by adding the contribution of individual loci (QTLs)
5. **Typically our models can only predict a small fraction of phenotypic variance – the so called "missing heritability" problem**

What causes missing heritability?

Possible explanations (non-exclusive):

- Incorrect heritability estimates
- Non-chromosomal elements
- Rare variants
- Structural variants
- Many common variants of low effect
- Epistasis

What causes missing heritability?

- Consider
  - f(ab) = 0
  - f(aB) = f(Ab) = 1
  - f(AB) = 0
- A and B will not be detected as QTLs as individually they have no effect on phenotype
- Assuming no environmental noise $H^2=1$ and $h^2=0$.
- Non-additive interactions can result from gene-gene interactions (epistasis)
  - Can be more than pairwise!
  - Considering all combinations of markers is in general not tractable because of multi-hypothesis limits
- Broad sense heritability includes additive genetic factors, dominance effects, gene-gene interactions, gene-environment interactions, non genomic inheritance

Remaining sources of heritability

- Gap between narrow- and broad-sense heritability implies genetic interactions
- For most traits, gaps not explained by found pairwise interactions
- Exception: maltose (71% of gap explained by one pairwise interaction)

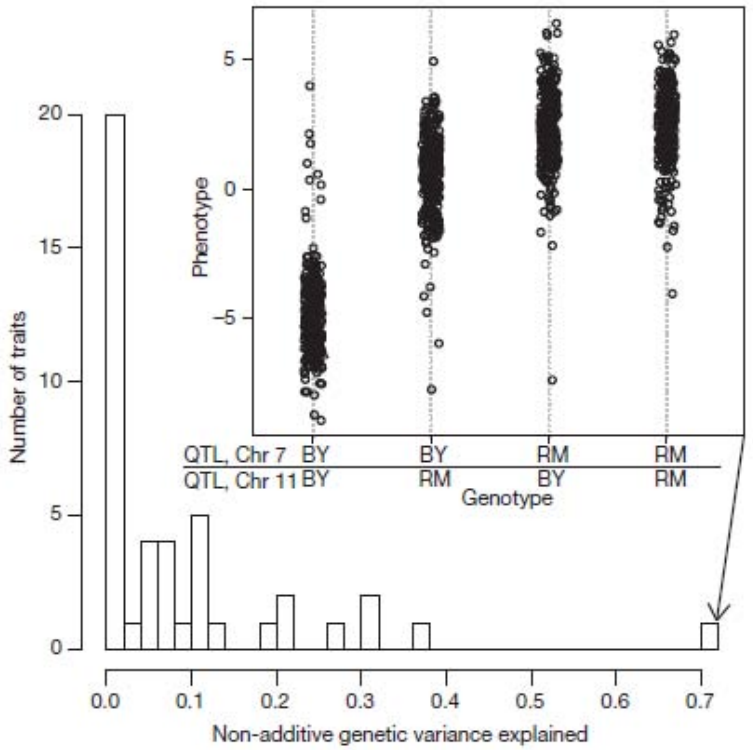
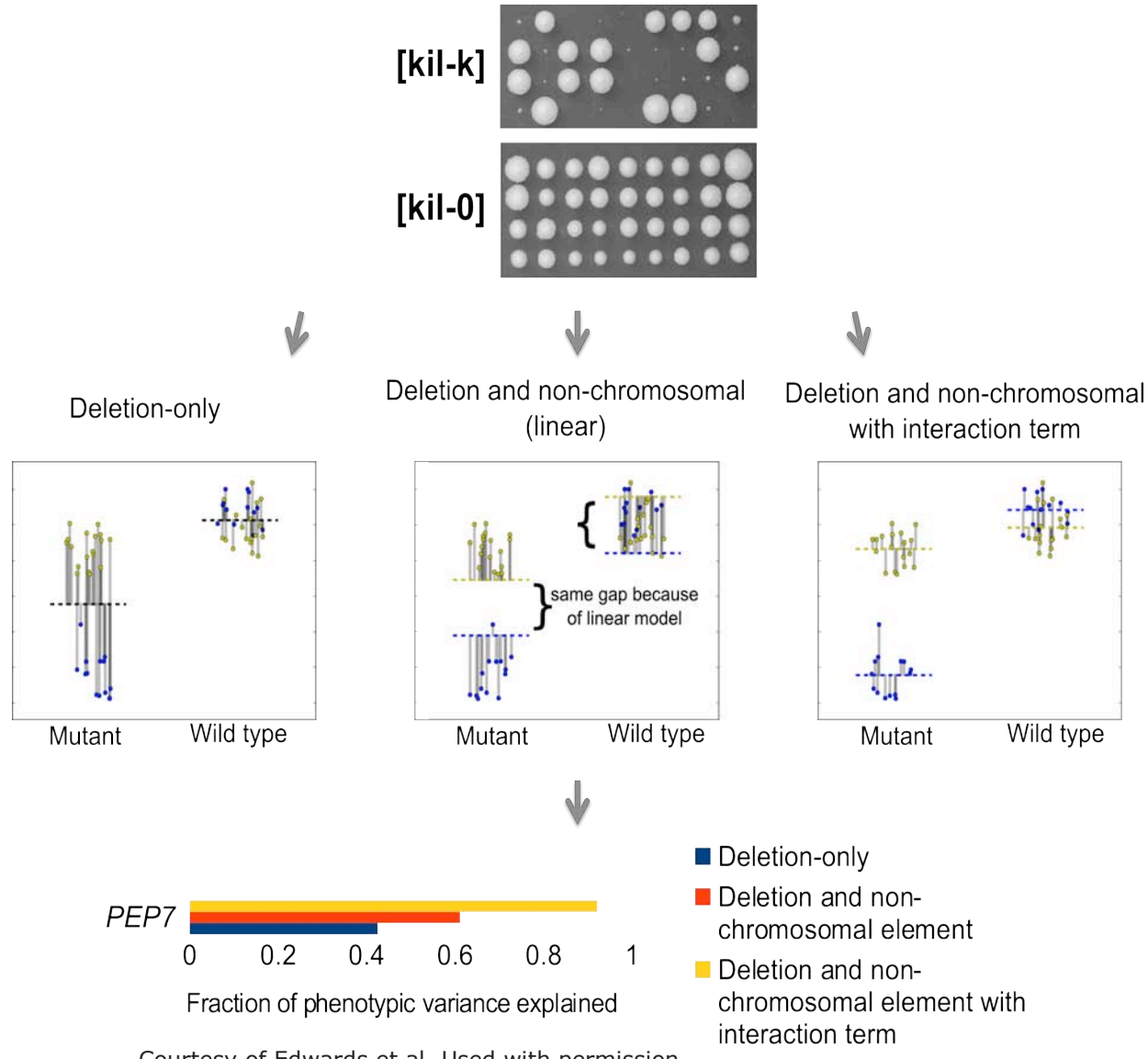


Figure 5 | Non-additive genetic variance explained by QTL–QTL interactions. A histogram of the fraction of non-additive genetic variance explained by detected QTL–QTL interactions per trait is plotted. The histogram is restricted to traits for which at least 10% of the total genetic variation is non-additive. Inset, phenotypes for growth in maltose are shown, grouped by two-locus genotypes at the two interacting QTL on chromosomes 7 and 11. This QTL–QTL interaction explained 71% of the difference between broad-sense and narrow-sense heritability.

Non-linear models reveal missing heritability from the interaction of chromosomal and non-chromosomal elements



Courtesy of Edwards et al. Used with permission.
Source: Edwards, Matthew D., Anna Symbor-Nagrabska, et al. "Interactions Between Chromosomal and Nonchromosomal Elements Reveal Missing Heritability." *Proceedings of the National Academy of Sciences* 111, no. 21 (2014): 7719-22.

Recent context

Problem: missing heritability for human diseases after hundreds of GWAS studies

Table 1 | Estimates of heritability and number of loci for several complex traits

Disease	Number of loci	Proportion of heritability explained
Age-related macular degeneration[72]	5	50%
Crohn's disease[21]	32	20%
Systemic lupus erythematosus[73]	6	15%
Type 2 diabetes[74]	18	6%
HDL cholesterol[75]	7	5.2%
Height[15]	40	5%
Early onset myocardial infarction[76]	9	2.8%
Fasting glucose[77]	4	1.5%

* Residual is after adjustment for age, gender, diabetes.

Source: Manolio, Teri A., Francis S. Collins, et al. "Finding the Missing Heritability of Complex Diseases." *Nature* 461, no. 7265 (2009): 747-53.

"Found" h^2 from Manolio et al. 2009

Discovering what is missing

- Use other data to determine relevance of markers (SNPs in enhancers, non-sense mutations, etc.) to reduce marker search space
- When relevant marker space is simplified can consider non-linear interactions
- Consider non-chromosomal genetic elements
- Use complementary data to determine marker interactions (protein-protein interaction data, etc.)
- Your research goes here!

FIN

7.91J / 20.490J / 20.390J / 7.36J / 6.802J / 6.874J / HST.506J Foundations of Computational and Systems Biology
Spring 2014