Phenotypic Variation and the Resemblance Between Relatives

Phenotype is the result of the interplay between genome and the environment

- All of these are phenotypes
    - Height
    - Amount of RNA transcribed from a given gene,
    - What you ate last Tuesday
- Nearly any phenotype represents the outcome of the information encoded by their genome played out through an incredibly complicated developmental, physiological and/or behavioral processes that in turn interact with a myriad of environmental and stochastic factors

The path from genotype through to phenotype

- Genotype-pheontype relationship
- **How phenotypic variation among individuals in a population arises as a result of genetic variation in the population**

- One simple way to measure this genotype-phenotype relationship is to calculate the phenotypic mean for each genotype at a locus.
- Example
  - flowering time gene (PtFT2) SNP
  - regression slope is –13.6
  - swapping a single T for a G allele moves the budset forward by 13.6 days, such that the GG homozygote is predicted to set buds 27.2 days earlier than the TT homozygote.
  - explains 62% of the variation in budset.

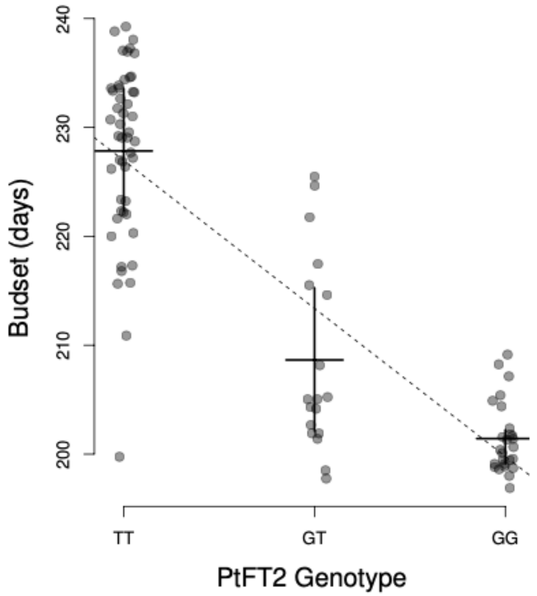


Figure 7.2: The effect of a flowering time gene (PtFT2) SNP on budset time in European aspen. Each dot gives the genotype-phenotype combination for an individual. The horizontal lines give the budset mean for each genotype and the vertical lines show the inter-quartile range. The dotted line gives the linear regression of phenotype on genotype. Thanks to Pär Ingvarsson for sharing these data from Wang et al. (2018).

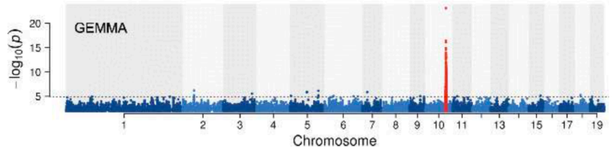


Figure 7.3: Manhattan plot of the p-value of the linear association between genotype and budset in aspen. Each dot represents the test at a single SNP, plotted at its physical coordinate in the genome. Different chromosomes are plotted in alternating colours. The SNPs surrounding the PtFT2 gene are shown in red. From Wang *et al.* (2018), licensed under CC BY 4.0.

The SNP with the most significant p-value is SNP in *PtFT2*. Note

Caveats

- Not an allele for budset, nor is PtFT2 a gene for budset.

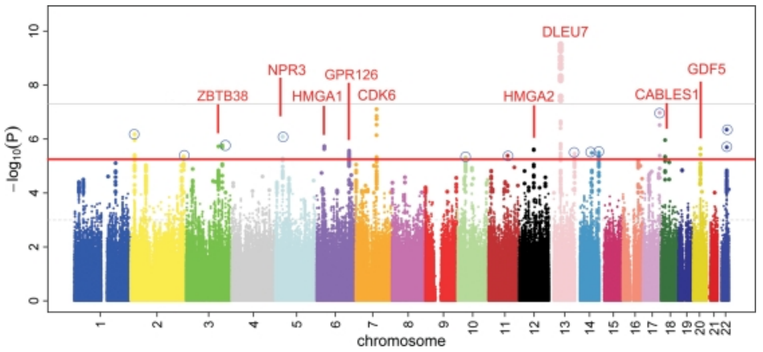
- allele is associated with budset in the sampled environments and populations.
- In a different set of environments, this allele's effects may be far smaller, and a different set of alleles may contribute to phenotype variation.
- PtFT2, the gene our focal SNP falls close to, is just one of many genes and molecular pathways involved in budset.
- A mutant screen for budset may uncover many genes with larger ef- fects; this gene is just a locus that happens to be polymorphic in this particular set of genotyped individuals.

Polygenic traits

- many phenotypes are likely much more genetically complex, involving the functional effect of many alleles at hundreds or thousands of polymorphic loci.

- Example: hundreds of small effect loci affecting human height have been mapped in Euro- pean populations to date

- Such genetically complex traits are called **polygenic traits**



A simple additive model of a trait

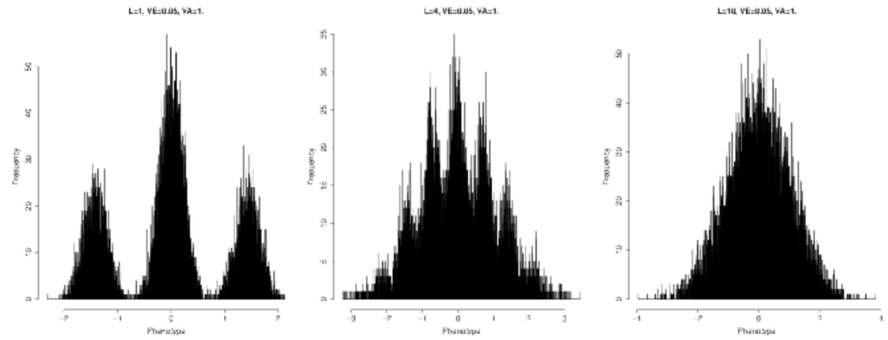
- X_i: phenotype of an individual
- $G_{i,l}$ = 0, 1, or 2, the number of copies of allele 1 she has at this SNP
- $X_{A,i}$ is the deviation away from the mean phenotype due to her genotype
- This X_E includes the systematic effects of the environment our individual finds herself in and all of the noise during development, growth, and the various random insults that life throws at our indi- vidual.

$$\mathbb{E}(X_i|G_{i,1},\cdots,G_{i,L}) = \mu + X_{A,i} = \mu + \sum_{l=1}^{L} G_{i,l}a_l$$

$$X_i = \mu + X_{A,i} + X_{E,i}$$

Why most polygenic traits are normally distributed

- If a reasonable number of loci contribute to variation in our trait then we can approximate the distribution of $X_{A,i}$ by a normal distribution due to the central limit theorem



Decomposition by parents

- X_{iM}: Contribution to phenotype by maternal alleles
- X_{iP}: Contribution to phenotype by paternal alleles

$$X_i = \mu + X_{iM} + X_{iP} + X_{iE}$$

Decomposition of total phenotypic variance

$$V_P = Var(X) = Var(X_A) + Var(X_E) = V_A + V_E$$

Our additive genetic variance can be written as

$$V_A = Var(X_A) = \sum_{l=1}^{L} Var(G_{i,l}a_l)$$

$$V_A = \sum_{l=1}^{L} a_l^2 2p_l(1-p_l)$$

- $Var(G_{i,l}a_l)$: the contribution of locus l to the additive variance among individuals.
- $2p_l(1-p_l)$ follows from the binomial sampling of two alleles per individual at each locus.

Decomposition of total phenotypic variance

- Similarly we can decompose the additive genetic variance V_A as the sum of maternal and paternal variance

$$V_A = Var(X_A) = Var(X_{M,i}) + Var(X_{P,i}) \quad (7.8)$$

$$Var(X_{M,i}) = Var(X_{M,i}) = {}^{V_A}/_2 \quad (7.9)$$

Polygenic score

- $$\mathbb{E}(X_i|G_{i,1},\cdots,G_{i,L}) = \mu + X_{A,i} = \mu + \sum_{l=1}^{L} G_{i,l}a_l$$
- a_l can be estimated from GWAS

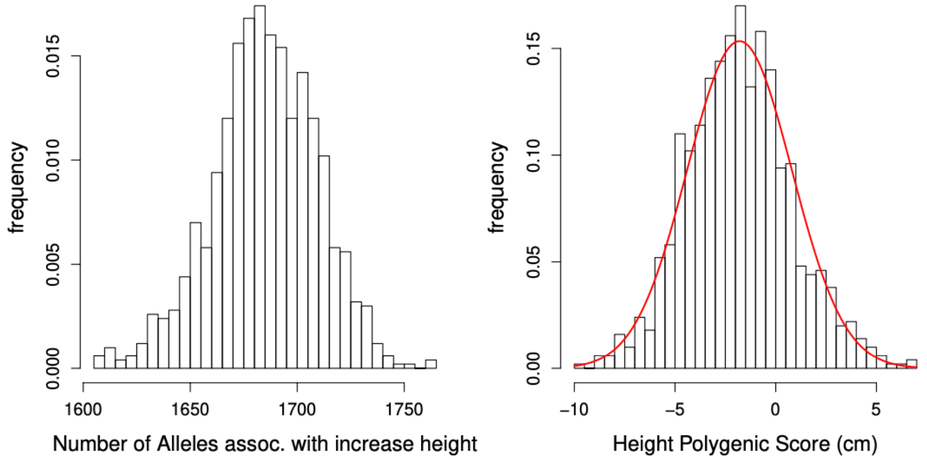


Figure 7.6: **Left)** The distribution of the number of height-increasing alleles that individuals carry at 1700 SNPs associated with height in the UK Biobank, for a sample of 1000 individuals. **right)** The distribution of the polygenic scores for these 1000 individuals. Plotted on top is a normal distribution with the same mean and variance. The empirical variance of these polygenic scores is 0.13, the additive genetic variance calculated by equation (7.7) is 0.135, so the two are in good agreement. Code here.

Heritability is key in helping us think about the evolution of phenotypes

- What proportion of the variation in our phenotype across individuals is due to genetic differences as opposed to environmental differences?

- If variation in our phenotype had no genetic basis, then no matter how much selection changes the mean phenotype within a generation the trait will not change over generations.

- Response to selection is proportional to h^2

$$h^2 = \frac{Var(X_A)}{V_P} = \frac{V_A}{V_P}$$

Heritability is a property of the population

- Heritability is a property of a sample from the population in a particular set of environments at a particular time.
- Changes in the environment may change the phenotypic variance.
- Changes in the environment may also change how our genetic alleles
- Example: h^2 of hair color may be very low in EA populations, but high in the global population

Narrow sense heritability vs. broad sense heritability

- Narrow sense heritability
- $$h^2 = \frac{Var(X_A)}{V_P} = \frac{V_A}{V_P}$$
- Broad sense heritability
- $H^2 = \frac{V_G}{V_P}$
- $V_G = V_A + V_D + V_{A\times A} + V_{A\times D} + V_{D\times D} + \cdots$

Estimate heritability - resemblance in parent-offspring

- At one locus:
  - M: mom; K: kid
  - $\text{Cov}(X_M, X_K) = \text{Cov}(\sum_i a_l G_{M,l}, \sum_i a_l G_{K,l}) = \sum_i \text{Cov}(a_l G_{M,l}, a_l G_{K,l})$
  - $\text{Cov}(a_l G_{M,l}, a_l G_{K,l}) = a_l^2 p_l (1 - p_l)$
  - $\text{Cov}(X_M, X_K) = \frac{1}{2} V_A$

$$V_A = \sum_{l=1}^{L} a_l^2 2 p_l (1 - p_l)$$

Estimate heritability - resemblance in midparent-offspring

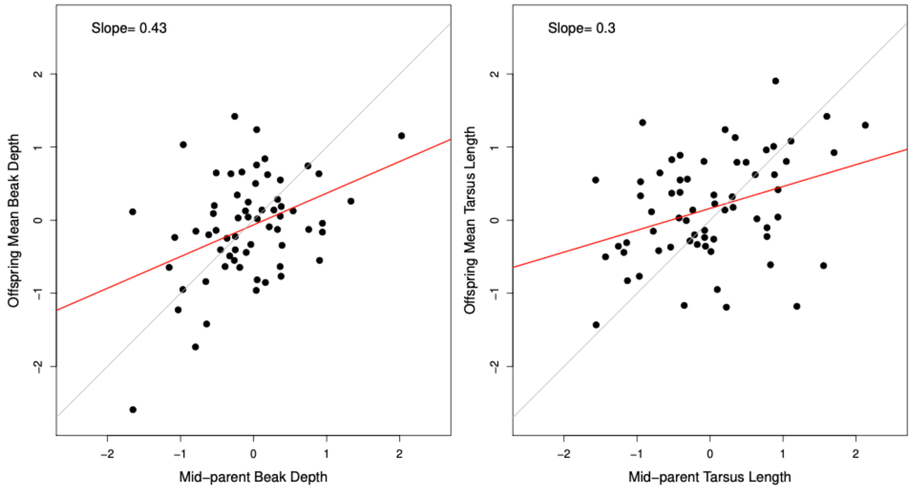
- At one locus:

$$\text{Cov}(\frac{1}{2}(X_M + X_P), X_K) = \frac{1}{2}\text{Cov}(\sum_i a_l G_{M,l}, \sum_i a_l G_{K,l})$$

- $$+\frac{1}{2}\text{Cov}(\sum_i a_l G_{P,l}, \sum_i a_l G_{K,l})$$

- $$\text{Cov}(\frac{1}{2}(X_M + X_P), X_K) = \frac{1}{2}V_A$$

- Least squares regression general solution:
- $X \sim Y$
- $y = \beta x + e$
- $\beta = \text{Cov}(X, Y)/\text{Var}(X)$
- $\beta_{\text{mid,kid}} = \dfrac{\text{Cov}(X_{\text{kid}}, X_{\text{mid}})}{\text{Var}(X_{\text{mid}})} = \dfrac{V_A}{V_P} = h^2$



The covariance between general pairs of relatives under an additive model

- $$Cov(X_1, X_2) = r_0 \times 0 + r_1 \frac{1}{2} V_A + r_2 V_A = 2F_{1,2} V_A$$

  - $F_{1,2}$ is our coefficient of kinship
  - i.e. the probability that two alleles sampled at random from our pair individuals 1 and 2 are IBD

“Animal” model for estimating heritability

- $$X_i = \mu + X_{A,i} + X_{E,i}$$
- $X_{E,i} \sim N(0, V_E)$
- $X_{A,i}$ is normally distributed across individuals
- $A_{ij} = 2F_{i,j}$
- $A_{ii} = 1.$
- Given the matrix A we can estimate V_A, using a linear mixed model

Estimating broad sense heritability

- Simplest method requires genetically identical replicates, i.e. twins, clones, etc.
  - $E(X_i, X_j) = V_A$
  - $H^2 = \text{corr}(X_i, X_j)$