Heritability Analyses

Readings:

- Manolio et al. 2009, Finding the missing heritability of complex diseases

- Lee SH et al. 2011, Estimating missing heritability for disease from genome-wide association studies.

Old Business

- Availability for make-up date for class:

  April 10 12-2pm?

  April 11 12-2pm?

Heritability Analyses

- Estimate phenotypic correlation between relatives
- Historically involves modeling phenotypic data from ***pedigrees*** without using genetic data
- Originally developed when genotyping (particularly genome-wide) was expensive, labor intensive, and not widely available
- Measure of the degree to which phenotypes are inherited (vs. due to environmental factors)

Quantitative Traits

- Historically defined in terms of quantitative traits (continuous phenotypes)
- Overall genetic component of trait relative to total observed phenotypic ***variation*** of the trait
- Quantitative traits (considered to be like common diseases) modeled as function of ***multiple*** QTLs (quantitative trait loci) and environmental components

General model (continuous phenotypes)

$$Y = \mu + \sum_{m=1,..,M} \{a_m X_m + d_m I\,[X_m = 1]\} + \epsilon$$

- Y (phenotype), M (unknown # of QTLs), X_m (# of disease alleles at the m^{th} locus)
- Parameters: μ (phenotypic mean in subjects without no copies of the QTL alleles), a_m is the additive component, d_m is the co-dominant component (allowing for departure from additive model), ε is environmental/non-genetic variability

Variance Components

$$\mathrm{Var}(Y) = \mathrm{Var}(G) + \mathrm{Var}(\epsilon) + 2\mathrm{Cov}(G, \epsilon)$$

$$\mathrm{Var}(G) = \mathrm{Var}(\sum_{m=1,..,M}(a_m X_m + d_m I[X_m = 1]))$$

- Usually assume covariance between is MUCH smaller than genetic contributions to variance (thus assume covariance of gene-environment interactions 0)

Broad-sense heritability

$$\mathrm{Var}(G)/\mathrm{Var}(Y)$$

- Proportion of overall phenotypic variation that can be attributed to genetic components
- Can also be divided into additive and co-dominant components…

Partitioning the variance

$$Var(G) = V_A + V_D$$

$$V_A = \sum_m 2p_m(1-p_m)(a_m + d_m(1-2p_m))^2$$

$$V_D = \sum_m (2p_m(1-p_m)d_m)^2$$

- Based on MAF (p); derived by conditioning on parental genotypes (mendelian inheritance)
- Additive component is the average effect of parental genotypes on offsprings phenotype (breeding coefficient)

Narrow-sense heritability

$$h^2 = V_A / \mathrm{Var}(Y)$$

- Except in situations where the mode of inheritance in thought to be heterozygous advantage (d>a) then d assumed << a and narrow-sense heritability thought to be a good estimate of heritability

- Advantage: can be estimated based on phenotypic family data

Derivation for phenotypes from trios

- Define Y_P as the average parent phenotype and Y_O as the offspring phenotype. Similarly X_P as the average parent genotype and X_O the offspring genotype.
- Assume genotype probability for parents follow independent binomial distribution with n=2 and p=MAF.
- Assume mendelian inheritance
- Assume 1 QTL and an additive model of inheritance

Derivation for phenotypes from trios…

- From model of Y defined on slide 5, the expected offspring phenotype and additive variance defined as:

$$E(Y_O) = E(Y_P) = E(aX_O) = \mu + 2ap$$

$$V_A = \text{var}(aX_O) = 2a^2p(1-p)$$

- Then the covariance between parent and offspring phenotypes is given as:

$$\text{Cov}(Y_O, Y_P) = a^2\text{cov}(X_O, X_P) = a^2[E(X_O X_P) - 4p^2]$$

Derivation for phenotypes from trios…

- Assuming Mendelian inheritance, you can show that:

$$\mathrm{cov}(X_O, X_P) = p(1-p)$$

$$\mathrm{Cov}(Y_O, Y_P) = a^2 p(1-p) = V_A/2$$

$$h^2 = 2\mathrm{Cov}(Y_O, Y_P)/\mathrm{Var}(Y) = 2\rho$$

- Such that narrow sense heritability equals 2x the correlation between offspring and average parental phenotype

Heritability estimation

- Estimators derived from sample variances and covariances; given in Falconer and Mackay (1996)
- Remember heritability depends on allele frequencies and environmental component both POPULATION dependent
- Also heritability can vary over time as allele frequencies can change (admixture) and environment can change
- Heritability does not depend on degree of relatedness of sample used for estimation (although based on correlation)

Continuous traits

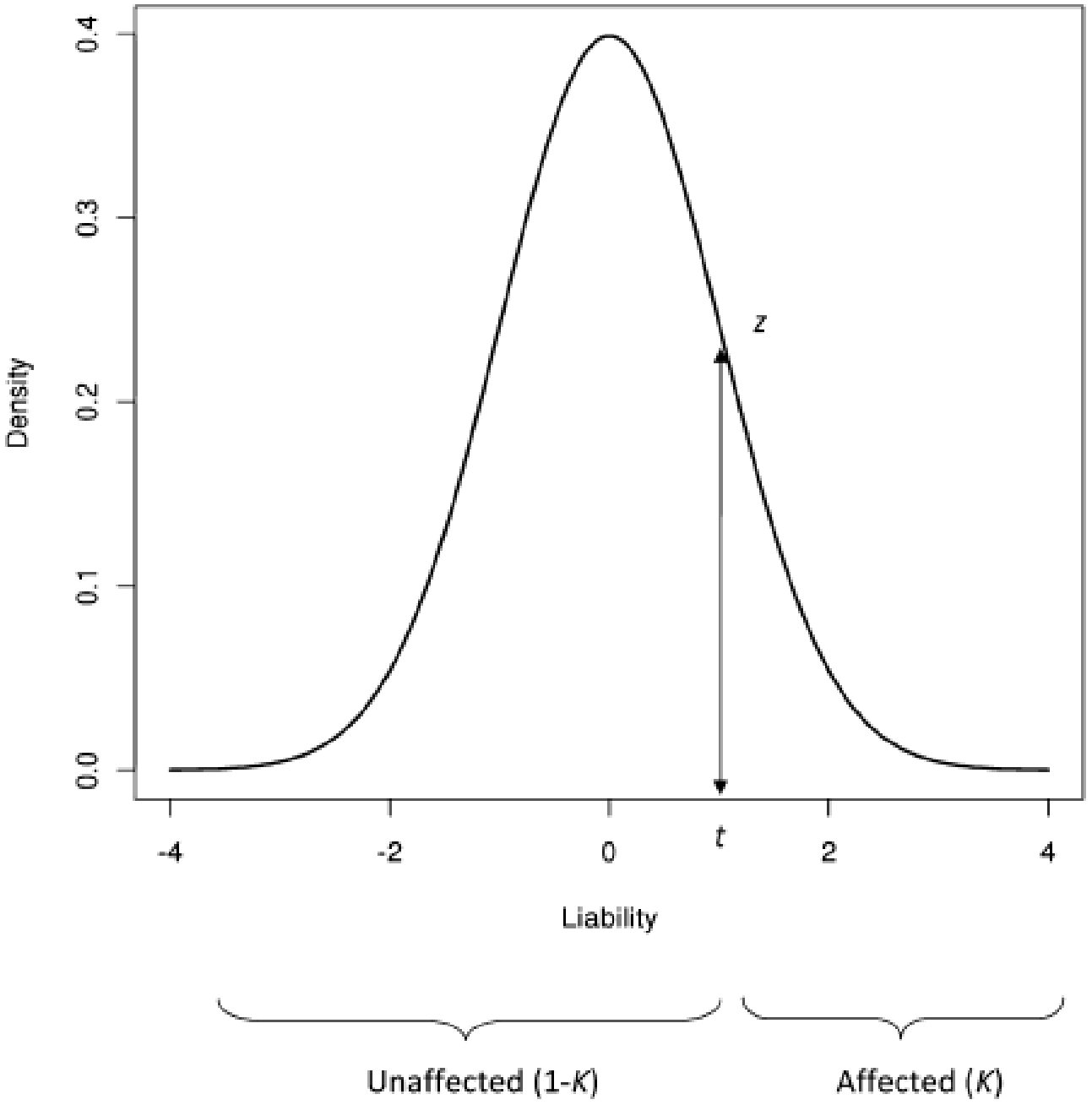
- Examples of phenotypes that show moderate/high genetic heritability include:
- Height: 80%
- Weight: 70-80%
- Cardiac ECG measurements: 25-50%
- Gene expression profiles: 25%-40%

Heritability in binary traits

- Binary traits: cases and controls
- Familial resemblance parameterized on unobserved continuous ***liability*** scale
- Heritability is independent of disease prevalence
- Use probit transformation to generate liability threshold model such that disease arises if the liability (normally distributed (0,1)) exceeds a certain threshold

Table 1. Recurrence risk (λ_R) to relatives (of type R) for several common complex genetic diseases ordered by prevalence (K)

Disease	Reference	K	λ_{MZ} [a]	λ_{Sib} [b]	λ_{OP}	H_{01}^2 [c] $= \frac{(\lambda_{MZ}-1)}{(1-K)}$	$\frac{(\lambda_{Sib}-1)}{(\lambda_{OP}-1)}$ [d]	$\frac{(\lambda_{MZ}-1)}{(\lambda_{Sib}-1)}$ [e]	$\frac{\lambda_{MZ}}{\lambda_{Sib}^2}$ [f]	h_L^2 [g]
Major depression (population cohort)	[27]	0.24	2	1.3		0.32		3.3	1.2	0.34
Age related macular degeneration	[28,29]	0.12	4.7	2.1		0.50		3.4	1.1	0.64
Myocardial infarction	[30]	0.056	4.6	3.2		0.21		1.6	0.4	0.72
Breast cancer	[31]	0.036	4.1	2.2	1.9	0.12	1.3	2.6	0.8	0.37
Type II diabetes	[32]	0.028	10.4	3.5		0.27		3.8	0.8	0.58
Asthma	[33]	0.019	6.6	3.4		0.11		2.3	0.6	0.49
Rheumatoid arthritis	[34]	0.01	12.2	3.6		0.11		4.3	0.9	0.42
Bipolar disorder	[5]	0.01	60	7	7	0.60	1.0	10	1.2	0.70
Schizophrenia	[3]	0.0085	52.1	8.6	10	0.44	0.8	6.7	0.7	0.76
Type I diabetes	[35]	0.005	79	14		0.39		6.0	0.4	0.85
Multiple sclerosis	[36]	0.001	190	20		0.19	~1	9.9	0.5	0.68
Crohn's disease	[37]	0.001	600	64		0.60		10	0.1	1.00
Ankylosis spondylitis	[6]	0.001	630	82	79	0.63	1.0	7.8	0.1	1.00
Systemic lupus erythematosus	[38]	0.001		29	27		1.1			0.80
	[39,40]	0.0003	774	65		0.24		12	0.2	0.84


Density
0.0
0.1
0.2
0.3
0.4
z
t
-4
-2
0
2
4
Liability
Unaffected (1-K)
Affected (K)

Liability threshold model

- Liability assumed to be the sum of environmental and additive genetic components from independent ***normal*** distributions

- Statistical methods developed for quantitative traits (estimation of heritability) can be applied to binary traits

$$\mathbf{l} = \mu \mathbf{1_N} + \mathbf{g} + \mathbf{e}$$

- Vector of liabilities distributed N(0,1) thus heritability on the liability scale: h^2 = Var(g)

Liability threshold model…

- Define g as probability of disease given genotype (x risk alleles out of 2n possible)

$$g_x = \Phi\left(\frac{u_x - t}{\sqrt{(1 - h_L^2)}}\right)$$

$$u_x = (x - 2np)a$$

Threshold and prevalence

- Threshold (t) defined such that the portion of the population the exceeds t is equal to the population prevalence K

- Derived from the inverse probability of the Z distribution:

$$t = \Phi^{-1}(1 - K)$$

$$\Phi(t) = 1 - K$$

- For example if K=5% then t=1.645

GWAS and heritability

- GWAS hypothesis: common disease-common variant theory
- Most common variants confer little incremental risk: odds ratios ranging from 1.1-1.5
- And in combination explain only a small fraction of the heritability of traits estimated from pedigrees
- Example: ~40 loci have been associated with height from GWAS but only explain 5% of the phenotypic variance!

Missing heritability

- Why is so much heritability unexplained by GWAS findings?
- Problem: substantial proportion of disease susceptibility conferred by genetic risk factors has not been identified
- Explanations include: larger number of variants of small effect (power too low), rare variants of larger effect (sequencing), structural variants, gene-gene interactions
- Structural: insertions, deletions, inversions, translocations

Heritability and prediction

- To know if risk variants together explain the total heritability in a population, use risk variants to predict phenotypes in a new set of subjects
- Correlate the predicted phenotypes with the observed phenotypes in the sample
- If correlation = estimated heritability then all the heritability is explained by the identified risk variants

Allelic architecture

- Defined as the #, type (structural etc), frequency, and effect size of risk variants
- Expected to differ across different diseases
- Example: few variants of large effect size explain much of the heritability of age-related macular degeneration (total: 50-70%) whereas many variants explain less of the total heritability of crohn's disease (total: 50-60%)

Allelic Architecture (# of SNPs)

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.

Rare variation

- Theoretically low frequency alleles with modest-large effect sizes could explain heritability of common diseases
- Example: only 20 variants of frequency 1% with allelic odds ratios of 3.0 could account for unexplained heritability in type II diabetes
- Few rare variants discovered to date either due to insufficient sample sizes or GWAS arrays do not comprehensively cover less common variation

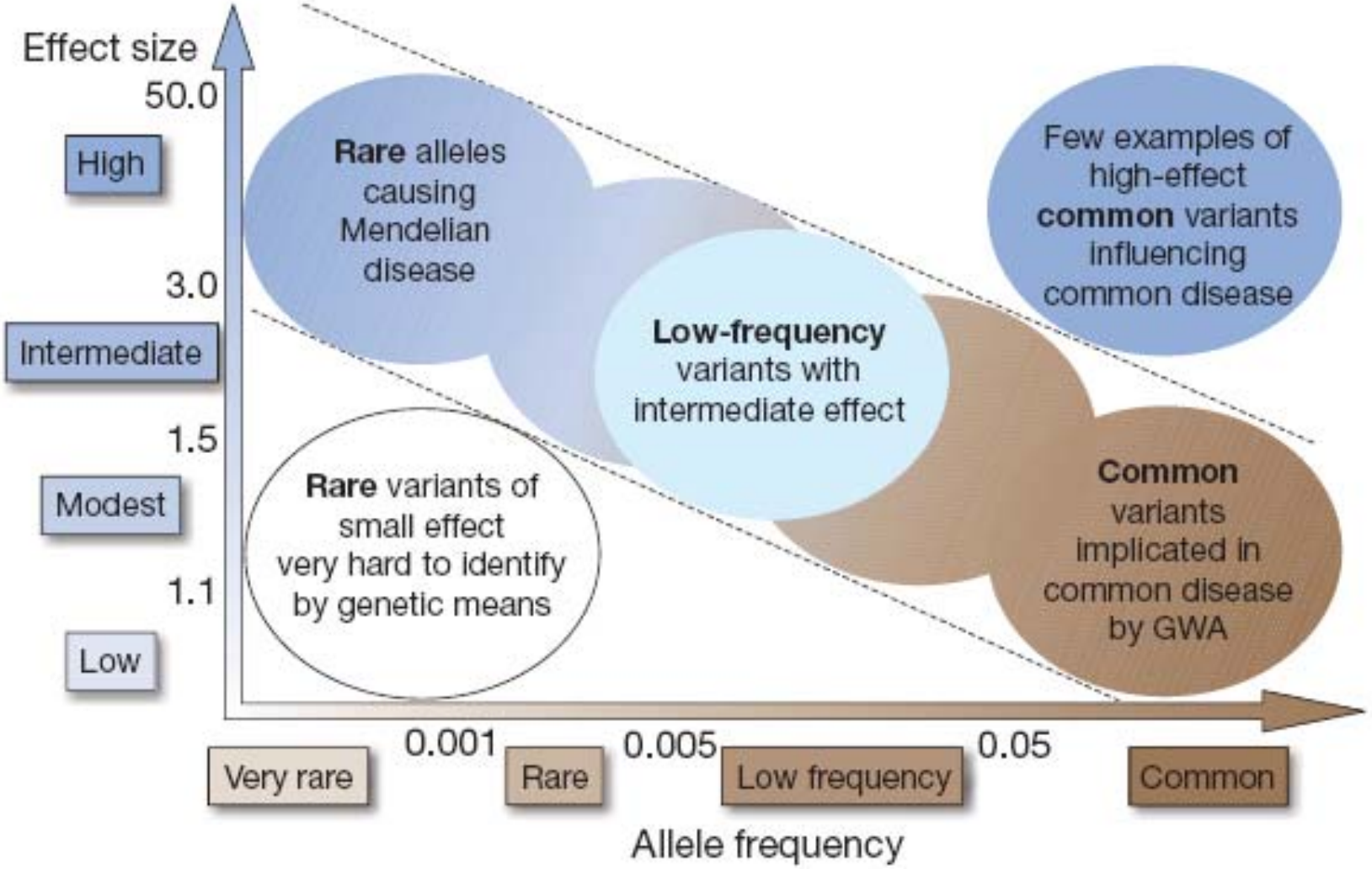


Figure 1 | Feasibility of identifying genetic variants by risk allele frequency and strength of genetic effect (odds ratio). Most emphasis and interest lies in identifying associations with characteristics shown within diagonal dotted lines. Adapted from ref. 42.

Next-gen sequencing

- Sample sizes required to detect associations increase linearly with 1/MAF
- Need LARGE samples to test associations with rare variations (however, small samples to identify variation)
- Study design: focus on extreme phenotypes, conduct studies in subjects of African descent (less LD), study families (rare variation over sampled, parent-of-origin effects), methods for testing pooled rare variation (by gene/region)

GCTA: Heritability estimates from GWAS

- Use GWAS data for continuous or binary phenotypes to relate phenotypic variance to estimates of IBD sharing between “unrelated individuals”
- Developed to identify “missing heritability” explained by SNPs on GWAS arrays or from imputation
- Estimates variance explained by all SNPs (+ imputed SNPs) rather than associations of individual SNPs with phenotypes

GCTA model

- Fit effects of all SNPs (g) as random effects by linear mixed model

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{g} + \boldsymbol{\varepsilon} \text{ with } \mathbf{V} = \mathbf{A}\sigma^2_{\mathrm{g}} + \mathbf{I}\sigma^2_{\varepsilon}$$

$$\mathbf{g} \sim N(0, \mathbf{A}\sigma^2_{\mathrm{g}})$$

- **A** is an nxn matrix called the genetic relationship matrix (GRM) estimated between individuals in the sample across all SNPs

Genetic relationship matrix

- Matrix of correlations between genotypes (assuming additive inheritance); here N is the total # of SNPs

$$A_{jk} = \frac{1}{N}\sum_{i=1}^{N}\frac{(x_{ij} - 2p_i)(x_{ik} - 2p_i)}{2p_i(1-p_i)}$$

$$\mathbf{g} \sim N(0, \mathbf{A}\sigma_g^2)$$

- Estimate σ_g^2 by restricted maximum likelihood (REML)

GCTA for binary traits: QC

- Experimental/genotyping artifacts unlikely to be correlated with continuous phenotypes
- Cases and controls are often collected/genotyped independently
- Without careful QC-ing genotyping differences between cases and controls will be attributed to missing heritability estimate

Liability scale

- Covariance between ***y*** (case/control) and ***l*** (liability) is equal to the height of the standard normal prob density function at threshold:

$$\mathrm{cov}(y,l) = E(y \cdot l) - E(y)E(l) = K1i + (1-K)0i_2 = Ki = z$$

$$\mathrm{E}(y \mid y > t) = \mathrm{i} = z/K$$

- Heritability on the observed scale is proportion of total variance (Bernoulli):

$$h_o^2 = \sigma_u^2 / [K(1-K)]$$

Liability scale and observed scale

- Relationship between heritability on the observed scale and the liability scale (after more algebra – see Lee et al.):

$$h_l^2 = h_o^2 K(1-K)/z^2$$

- Valid only in samples without ascertainment! So we have to adjust for inflated proportion of cases…

Ascertainment adjustment

- E(y)=P (proportion of cases in the sample) which gives:

$$\mathrm{var}(y_{cc}) = P(1-P)$$

$$\mathrm{var}(l_{cc}) = \sigma^2_{l_{cc}} \quad = 1 + i\lambda(t - i\lambda)$$

$$\lambda = (P-K)/(1-K)$$

- See Lee et al. for derivation …

$$h^2_l = \sigma^2_g = \widehat{h}^2_{o_{cc}} \frac{K(1-K)}{z^2} \frac{K(1-K)}{P(1-P)}$$

WTCCC type I diabetes heritability

Table 5. Estimated Genetic Variance on the Observed and Liability Scale Explained by All SNPs for Type I Diabetes in WTCCC Data

Threshold[a]	No. SNP[b]	Estimate[c] (SE)	LR	Adjusted[d] (SE)	Transformed[e] (SE)
MAF > 0.01					
200	318,044	0.57 (0.07)	70.36	0.65 (0.08)	0.32 (0.04)
20	289,463	0.56 (0.07)	70.32	0.65 (0.08)	0.32 (0.04)
7	238,805	0.52 (0.07)	61.51	0.61 (0.08)	0.30 (0.04)
4	178,892	0.51 (0.07)	64.74	0.64 (0.08)	0.31 (0.04)
MAF > 0.05					
200	289,693	0.54 (0.07)	70.48	0.61 (0.08)	0.30 (0.04)
20	262,091	0.53 (0.07)	70.49	0.61 (0.08)	0.30 (0.04)
7	216,136	0.49 (0.06)	61.81	0.57 (0.08)	0.28 (0.04)
4	162,162	0.48 (0.06)	63.54	0.58 (0.08)	0.29 (0.04)

WTCCC type I diabetes heritability

Table 6. Estimated Genetic Variance on the Observed and Liability Scale Explained by All SNPs for Type I Diabetes from an Analysis without Chromosome 6 or of Chromosome 6 Only

Threshold[a]	No. SNP[b]	Estimate[c] (SE)	LR	Adjusted[d] (SE)	Transformed[e] (SE)
Analysis without chromosome 6					
200	297,028	0.23 (0.07)	11.98	0.26 (0.08)	0.13 (0.04)
20	270,332	0.22 (0.07)	10.66	0.25 (0.08)	0.12 (0.04)
7	223,039	0.20 (0.07)	9.08	0.23 (0.08)	0.12 (0.04)
4	167,099	0.20 (0.06)	10.17	0.26 (0.08)	0.13 (0.04)
Analysis of chromosome 6 only					
200	21,016	0.33 (0.02)	268.55	0.37 (0.03)	0.18 (0.01)
20	19,131	0.33 (0.02)	278.09	0.37 (0.03)	0.18 (0.01)
7	15,766	0.32 (0.02)	255.65	0.36 (0.03)	0.18 (0.01)
4	11,793	0.31 (0.02)	264.63	0.38 (0.03)	0.19 (0.01)

WTCCC bipolar disorder heritability

Table 4. Estimated Genetic Variance on the Observed and Liability Scale Explained by All SNPs for Bipolar Disorder in WTCCC Data

Threshold[a]	No. SNP[b]	Estimate[c] (SE)	LR	Adjusted[d] (SE)	Transformed[e] (SE)
MAF > 0.01					
200	321605	0.71 (0.07)	107.76	0.81 (0.08)	0.41 (0.04)
20	291724	0.68 (0.07)	100.48	0.78 (0.08)	0.40 (0.04)
7	245127	0.65 (0.07)	94.69	0.76 (0.08)	0.38 (0.04)
4	187597	0.62 (0.07)	92.21	0.76 (0.08)	0.38 (0.04)
MAF > 0.05					
200	292969	0.68 (0.07)	110.45	0.77 (0.08)	0.39 (0.04)
20	264151	0.65 (0.07)	103.46	0.75 (0.08)	0.38 (0.04)
7	221947	0.62 (0.07)	97.64	0.72 (0.08)	0.37 (0.04)
4	170143	0.60 (0.06)	95.47	0.73 (0.08)	0.37 (0.04)

Downloading GCTA

- Out of date version (1.04) for MAC or PC no longer available on the web
- Download here: gcta_1.04.zip (necessary for Problem Set #4)
- See documentation: www.complextraitgenomics.com/software/gcta/index.html

Next class: Pathway analysis/eQTLs

Readings:

- Nicolae et al. 2010 Trait-associated SNPs are more likely to be eQTLs: annotation to enhance discovery from GWAS.

- De la Cruz et al. 2010 Gene, region, and pathway level analyses in whole-genome studies