Use of R in Genetic Epidemiology Designs -Power/Sample Size Considerations

Jing Hua Zhao [1,2]

[1]MRC Epidemiology Unit
[2]Institute of Metabolic Science
Addenbrooke's Hospital
Cambridge CB2 0QQ
United Kingdom
http://www.mrc-epid.cam.ac.uk/~jinghua.zhao
E-mail: jinghua.zhao@mrc-epid.cam.ac.uk

July 21, 2010

Table of contents

Some terminology

- Genes, Chromosome, markers
- Alleles, genotypes, haplotypes
- Phenotypes, mode of inheritance, penetrance
- Mendelian laws of inheritance
- Hardy-Weinberg equilibrium
- linkage disequilibrium
- Gene-environment interaction

Genetic epidemiology

- It is the study of the role of genetic factors in determining health and disease in families and in populations, and the interplay of such genetic factors with environmental factors, or "a science which deals with the aetiology, distribution, and control of diseases in groups of relatives and with inherited causes of disease in populations" (http://en.wikipedia.org).
- It customarily includes study of familial aggregation, segregation, linkage and association. It is closely associated with the development of statistical methods for human genetics which deals with these four questions. The last two questions can only be answered if appropriate genetic markers available (Elston & Ann Spence. Stat Med 2006;25:3049-80).

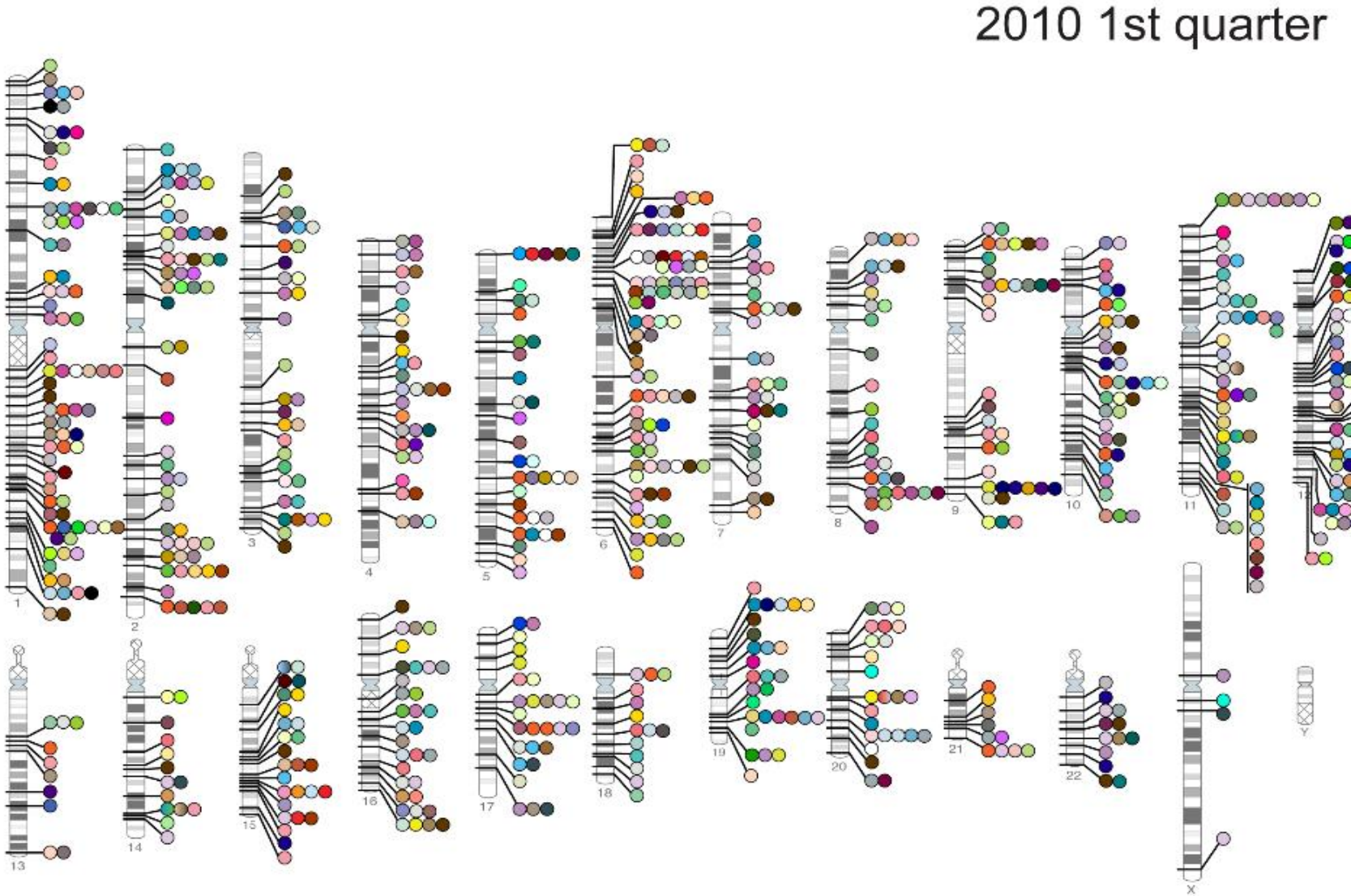
Linkage studies

- It is the study of cosegregation between genetic markers and putative disease loci, and has been very successful in localizing rare, Mendelian disorders but since has difficulty for traits which do not strictly follow Mendelian mode of inheritance, considerable linkage heterogeneity and it has limited resolution.
- It typically involves parametric (model-based) and nonparametric (model-free) methods, the latter most commonly refers to allele-sharing methods.
- The underlying concepts are nevertheless very important. It can still be useful in providing candidates for fine-mapping and association studies.
- With availability of whole genome data, it is possible to infer relationship or correlation between any individuals in a population.

Association studies

- They focus on association between particular allele and trait; it is only feasible with availability of dense markers.
- It has traditionally applied to both relatives in families and population sample. For the latter there has been serious concern over spurious association due to difference in allele frequencies between hidden sub-populations in a sample.
- A range of considerations has been made (Balding. Nat Rev Genet 2006;7:781-91) but the availability of whole genome data again refresh views including statistical examination of population substructure.

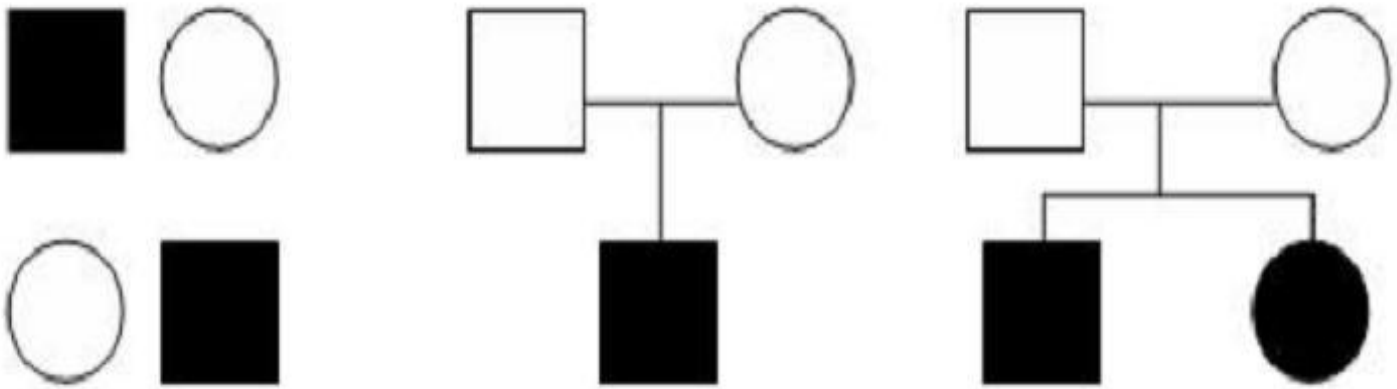
Published genome-wide associations



Annotations (http://www.genome.gov/GWAStudies)

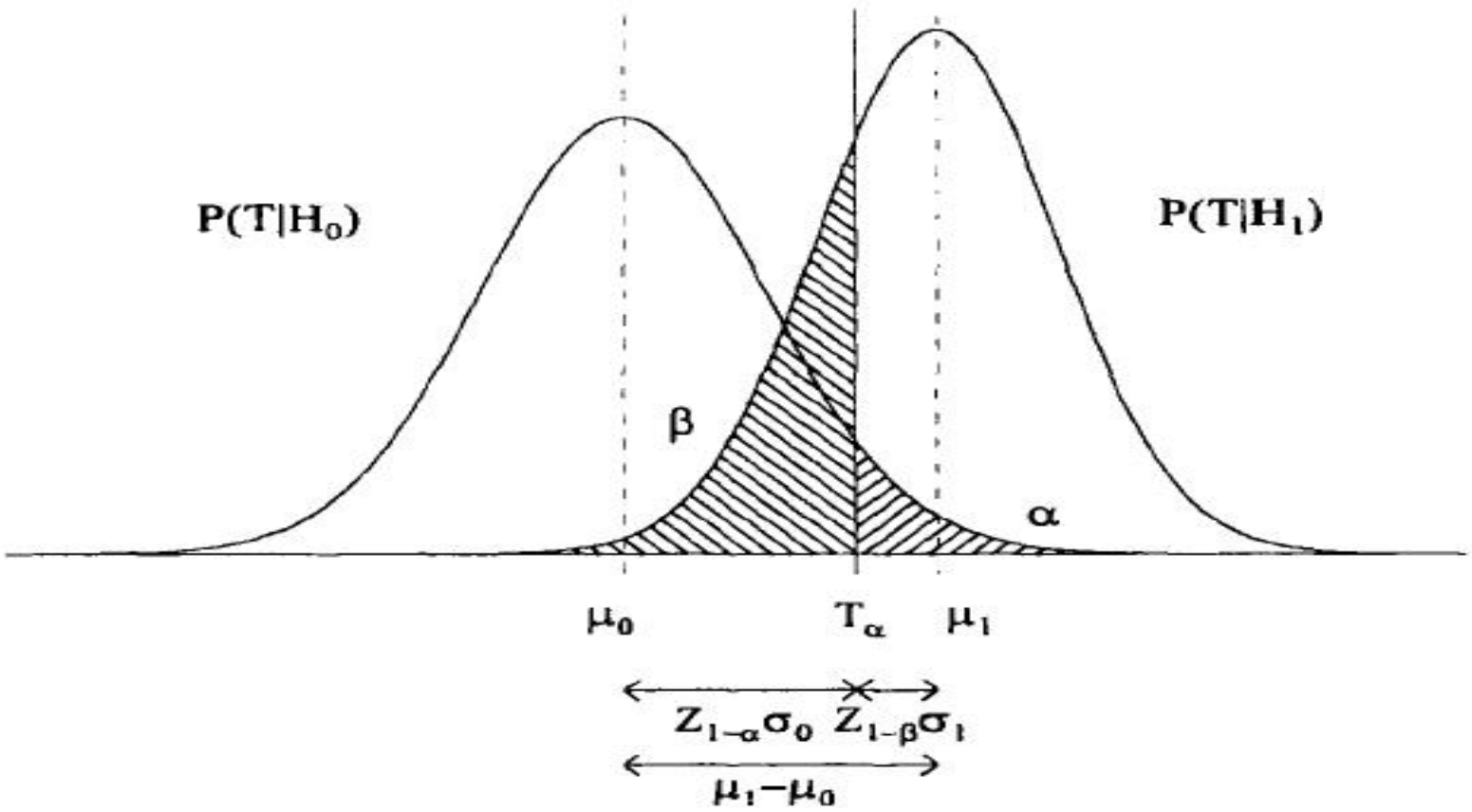
Major study designs

Three common genetic association designs involving unrelated individuals (left), nuclear families with affected singletons (middle) and affected sib-pairs (right). Males and females are denoted by squares and circles with affected individuals filled with black colors and unaffected individuals being empty



See Risch N, Merikangas K. Science 1996; 273:1516-7 and Thomas D. Nat Rev Genet 2010; 11:259-72.

A conceptual picture based on a test of $H_0 : \mu = \mu_0$ vs $H_1 : \mu = \mu_1 > \mu_0$ from a normal distribution



Sample size calculation based on normal distribution

Let $T \sim N(\mu_1, \sigma_1^2)$, we have the following steps,

- $Z = \frac{T-\mu_0}{\sigma_0} \sim N(\frac{\mu_1-\mu_0}{\sigma_0}, \frac{\sigma_1^2}{\sigma_0^2})$.
- $\beta = P(Z < Z_{1-\alpha}|\mu_1, \sigma_1^2) = \Phi\left(\frac{Z_{1-\alpha} - \frac{\mu_1-\mu_0}{\sigma_0}}{\frac{\sigma_1}{\sigma_0}}\right)$ and $Z_\beta = \frac{Z_{1-\alpha}\sigma_0-(\mu_1-\mu_0)}{\sigma_1}$.
- Since $Z_\beta = -Z_{1-\beta}$ and we are interested in $1-\beta$, $Z_{1-\beta} = \frac{(\mu_1-\mu_0)-Z_{1-\alpha}\sigma_0}{\sigma_1}$, $|\mu_1 - \mu_0| = Z_{1-\alpha}\sigma_0 + Z_{1-\beta}\sigma_1$.
- As $\sigma_i \equiv \sigma_i/N$, $i = 1, 2$. $\sqrt{N}|\mu_1 - \mu_0| = Z_{1-\alpha}\sigma_0 + Z_{1-\beta}\sigma_1$.

$$N = \left(\frac{Z_{1-\alpha}\sigma_0 + Z_{1-\beta}\sigma_1}{\mu_1 - \mu_0}\right)^2$$

Sample size estimation for affected sib-pair linkage and association

The mean and variance for the designs considered above were given in Risch & Merikangas (1996) and described in Zhao J Stat Soft 2007; 23(8):1-8, both under multiplicative model.

Let γ=genotypic risk ratio; p=frequency of disease allele A; Y=probability of allele sharing; N_L=number of ASP families required for linkage; P_A=probability of transmitting disease allele A; H_1, H_2=proportions of heterozygous parents; N_{tdt}=number of family trios; N^*_{asp}=number of ASP. families

The following tables were based on refined pbsize and fbsize functions in R/gap.

Power of linkage versus association

γ	p	Y	N_{asp}	P_A	H_1	N_{tdt}	H_2	N^*_{asp}	λ_o/λ_s
4.0	0.01	0.52	6402	0.80	0.05	1201	0.11	257	1.08/1.09
4.0	0.10	0.60	277	0.80	0.35	165	0.54	53	1.48/1.54
4.0	0.50	0.58	446	0.80	0.50	113	0.42	67	1.36/1.39
4.0	0.80	0.53	3024	0.80	0.24	244	0.16	177	1.12/1.13
2.0	0.01	0.50	445964	0.67	0.03	6371	0.04	2155	1.01/1.01
2.0	0.10	0.52	8087	0.67	0.25	761	0.32	290	1.07/1.08
2.0	0.50	0.53	3753	0.67	0.50	373	0.47	197	1.11/1.11
2.0	0.80	0.51	17909	0.67	0.27	701	0.22	431	1.05/1.05
1.5	0.01	0.50	6944779	0.60	0.02	21138	0.03	8508	1.00/1.00
1.5	0.10	0.51	101926	0.60	0.21	2427	0.25	1030	1.02/1.02
1.5	0.50	0.51	27048	0.60	0.50	1039	0.49	530	1.04/1.04
1.5	0.80	0.51	101926	0.60	0.29	1820	0.25	1030	1.02/1.02

Sample sizes required for association detection using population data with given prevalences

γ	p	1%	5%	10%	20%
4.0	0.01	46681	8959	4244	1887
4.0	0.10	8180	1570	744	331
4.0	0.50	10891	2091	991	441
4.0	0.80	31473	6041	2862	1272
2.0	0.01	403970	77530	36725	16323
2.0	0.10	52709	10116	4792	2130
2.0	0.50	35285	6772	3208	1426
2.0	0.80	79391	15237	7218	3208
1.5	0.01	1599920	307056	145448	64644
1.5	0.10	192105	36869	17465	7762
1.5	0.50	98013	18811	8911	3961
1.5	0.80	192105	36869	17465	7762

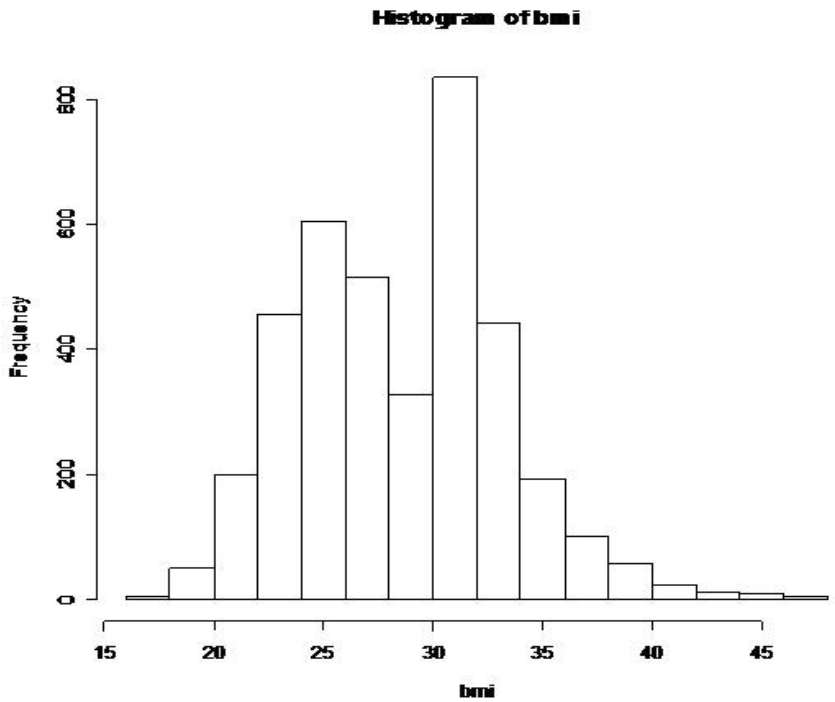
EPIC-Norfolk obesity project

- The initial design of the study was case-control (e.g., WTCCC with seven cases and controls) with 3425 cases and 3400 controls.
  - It is potentially more powerful.
  - Controls are selected, however.
- It has therefore been changed into case-cohort design, in which cases are defined to be individuals whose BMI above 30 and controls are a random sample (sub-cohort) of the EPIC-Norfolk cohort which includes obese individuals.
- The sub-cohort is representative of the whole population and allows for a range of traits to be examined.
- There is more work to do with two-stage design.
- The problem of Mendelian randomisation can be considered in a general framework.

Power/sample size

- It started with assessment of how the power is compromised relative to the original case-control design.
- This was followed by power/sample size calculation using methods established by Cai & Zeng (Biometrics 2004, 60:1015-1024) as implemented in the R/gap function ccsize, noting a number of assumptions.
- More practically, we took the subcohort sample size to be 2,500, i.e., 10% of a total of 25,000 individuals as a rough representative sample.

Case-cohort sample distribution



The case-cohort sample of the EPIC-Norfolk obesity genetics project is a combination of the sub-cohort sample and case sample which is truncated from the whole cohort at BMI=30

N=25,000, $\alpha = 5 \times 10^{-8}$ (p_D=prevalence, p_1=frequency, hr=hazard ratio)

pD	p1	hr	ssize	pD	p1	hr	ssize
0.3	0.1	1.3	14391	0.2	0.2	1.4	3164
0.3	0.1	1.4	5732	0.2	0.3	1.3	4548
0.3	0.2	1.2	21529	0.2	0.3	1.4	2152
0.3	0.2	1.3	5099	0.2	0.4	1.2	20131
0.3	0.2	1.4	2613	0.2	0.4	1.3	3648
0.3	0.3	1.2	11095	0.2	0.4	1.4	1805
0.3	0.3	1.3	3490	0.2	0.5	1.2	17120
0.3	0.3	1.4	1882	0.2	0.5	1.3	3422
0.3	0.4	1.2	8596	0.2	0.5	1.4	1713
0.3	0.4	1.3	2934	0.1	0.2	1.4	8615
0.3	0.4	1.4	1611	0.1	0.3	1.4	3776
0.3	0.5	1.2	7995	0.1	0.4	1.3	13479
0.3	0.5	1.3	2786	0.1	0.4	1.4	2824
0.3	0.5	1.4	1538	0.1	0.5	1.3	10837
0.2	0.1	1.4	9277	0.1	0.5	1.4	2606
0.2	0.2	1.3	7725				

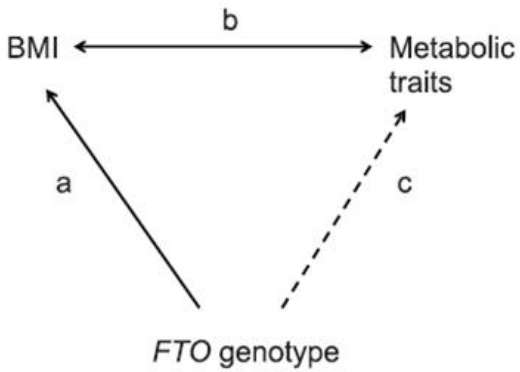
Does it work? The LDL example (Sandhu et al. Lancet 2008, 371:483-91)

	Study 1 (EPIC-Norfolk subcohort) n=2269		Study 2 (EPIC-Norfolk obese set) n=1009		Study 3 (1958 British birth cohort) n=1375		Study 4 (CoLaus) n=5367		Study 5 (GEMS study) n=1665	
	β coeff (SE)	p value	β coeff (SE)	p value	β coeff (SE)	p value	β coeff (SE)	p value	β coeff (SE)	p value
rs4420638	0·24 (0·04)	$1·9\times10^{-9}$	0·14 (0·06)	0·02	0·25 (0·04)	$2·8\times10^{-9}$	0·05 (0·01)	$6·2\times10^{-12}$	0·04 (0·01)	$5·6\times10^{-3}$
rs599839	−0·15 (0·04)	$5·8\times10^{-5}$	−0·23 (0·06)	$7·6\times10^{-5}$	−0·14 (0·04)	$4·3\times10^{-4}$	−0·04 (0·01)	$1·6\times10^{-7}$	−0·06 (0·01)	$2·0\times10^{-5}$
rs4970834	−0·13 (0·04)	$1·1\times10^{-3}$	−0·18 (0·06)	$5·5\times10^{-3}$	−0·11 (0·04)	0·01	−0·04 (0·01)	$1·9\times10^{-6}$	−0·04 (0·01)	$2·8\times10^{-3}$
rs562338	−0·17 (0·04)	$6·0\times10^{-6}$	−0·11 (0·06)	0·07	−0·18 (0·05)	$1·1\times10^{-4}$	−0·03 (0·01)	$2·7\times10^{-6}$	−0·02 (0·01)	0·18
rs7575840	0·15 (0·03)	$6·3\times10^{-6}$	0·15 (0·05)	$2·4\times10^{-3}$	0·04 (0·04)	0·26	0·03 (0·01)	$1·9\times10^{-6}$	0·02 (0·01)	0·13
rs478442	−0·16 (0·04)	$2·1\times10^{-5}$	−0·07 (0·06)	0·25	−0·16 (0·04)	$3·6\times10^{-4}$	−0·03 (0·01)	$2·7\times10^{-5}$	−0·02 (0·01)	0·06
rs4591370	−0·17 (0·04)	$7·7\times10^{-6}$	−0·06 (0·06)	0·28	−0·16 (0·04)	$4·2\times10^{-4}$	−0·03 (0·01)	$3·2\times10^{-5}$	−0·02 (0·01)	0·06
rs4560142	−0·16 (0·04)	$1·6\times10^{-5}$	−0·06 (0·06)	0·27	−0·16 (0·04)	$4·2\times10^{-4}$	−0·03 (0·01)	$3·5\times10^{-5}$	−0·03 (0·01)	0·05
rs576203	−0·16 (0·04)	$1·2\times10^{-5}$	−0·07 (0·06)	0·25	−0·16 (0·04)	$3·5\times10^{-4}$	−0·03 (0·01)	$3·5\times10^{-5}$	−0·02 (0·01)	0·06
rs506585	−0·16 (0·04)	$1·7\times10^{-5}$	−0·06 (0·06)	0·31	−0·16 (0·04)	$3·5\times10^{-4}$	−0·03 (0·01)	$4·2\times10^{-5}$	−0·03 (0·01)	0·05
rs488507	−0·14 (0·04)	$1·3\times10^{-4}$	−0·07 (0·06)	0·25	−0·16 (0·04)	$3·3\times10^{-4}$	−0·03 (0·01)	$3·4\times10^{-5}$	−0·02 (0·01)	0·07
rs538928	−0·16 (0·04)	$5·0\times10^{-5}$	−0·01 (0·06)	0·92	−0·16 (0·04)	$3·5\times10^{-4}$	−0·03 (0·01)	$3·6\times10^{-5}$	−0·02 (0·01)	0·05
rs10402271	0·04 (0·03)	0·17	0·11 (0·05)	0·02	0·12 (0·04)	$7·5\times10^{-4}$	0·02 (0·01)	$5·2\times10^{-4}$	0·04 (0·01)	$8·3\times10^{-4}$
rs693	−0·12 (0·03)	$1·3\times10^{-4}$	−0·07 (0·05)	0·15	−0·06 (0·03)	0·06	−0·03 (0·01)	$1·0\times10^{-5}$	−0·02 (0·01)	0·16

Webtable 3: Associations between Affymetrix SNPs with a combined p value of $<1·0\times10^{-7}$ and circulating concentrations of LDL cholesterol in independent study populations

FTO-BMI-metabolic traits association

In analogy to the *FTO*-T2D association mediated by BMI, metabolic traits have been considered (Freathy et al. Diabetes 2008, 57:1419-26) in a so-called Mendelian randomisation study of causal association.

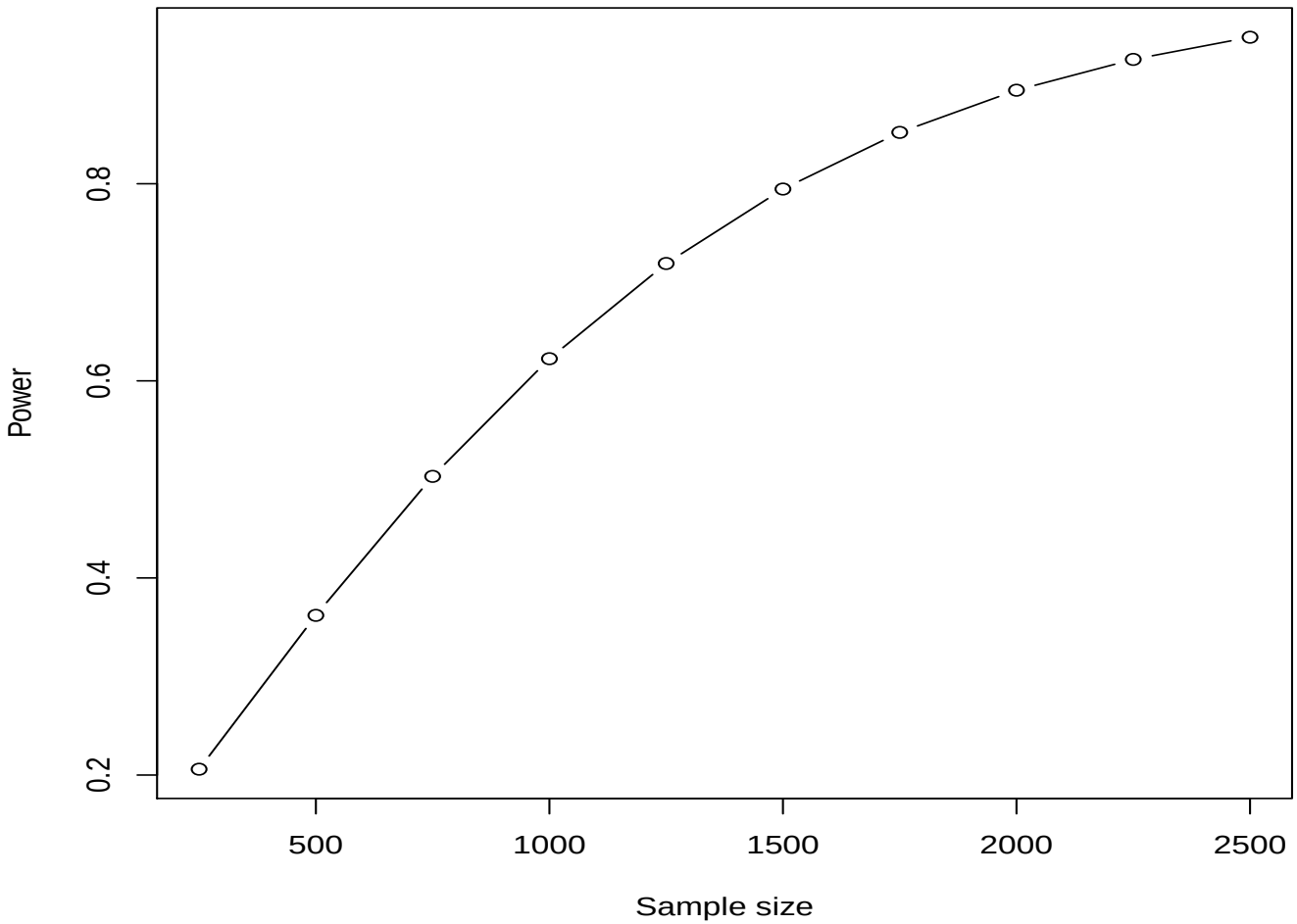


There are *FTO*-BMI (a) and BMI-metabolic traits (b) associations, and BMI is a mediator between *FTO* and metabolic traits ($c = a \times b$).

Power based on ab

- We can of course perform simulations to obtain power estimate but it would be somewhat involved.
- Instead, standard error of *FTO*-BMI-T2D can be calculated which can form the basis of power calculation (Kline RB. Principles and Practice of Structural Equation Modeling, Second Edition, The Guilford Press 2005).
- We implement this in ab function in R/gap.
- We have for EPIC-Norfolk 25,000, SNP-BMI regression coefficient (SE) of 0.15 (0.01), and BMI-T2D log(1.19) (0.01). We consider $\alpha = 0.05$.
- Criticism arisen from this posthoc power calculation could be alleviated when we allow for a range of sample sizes to be considered in the next slide.

SNP-BMI-T2D in EPIC-Norfolk study



How have we obtained the unexpected with T2D?

- We have customarily used $T2D = i_1 + cSNP + e_1$, $T2D = i_2 + c^{'}SNP + bBMI + e_2$, $BMI = i_3 + aSNP + e_3$; $\hat{a}\hat{b}$ is called the "expected" and $\hat{c} - \hat{c}'$ the "observed" which does not change even though the size of the mediated effect increases, i.e., $\hat{\sigma}^2_{T2D} = \hat{c}^2\hat{\sigma}^2_{SNP} + \frac{\pi^2}{3}$, $\hat{\sigma}^2_{T2D} = \hat{c}'^2\sigma^2_{SNP} + \hat{b}^2\hat{\sigma}^2_{BMI} + 2\hat{c}'\hat{b}\hat{\sigma}_{SNPBMI} + \frac{\pi^2}{3}$.
- It is recommended that $\hat{c}_{corrected} = \hat{c}\sqrt{1 + \frac{\hat{b}^2\hat{\sigma}^2_{SNPBMI}}{\frac{\pi^2}{3}}}$ be used instead.
- We replace $\frac{\pi^2}{3}$ with 1 for probit regression.

Generic power calculation for mediation analysis

- One may need to consider a range of outcomes such as binary, continuous, count, time-to-event and longitudinal data.
- A unified frame work was discussed by Vittinghoff E, et al. (2009) Stat Med 28:541-57. The model takes into account binary or continuous primary and mediation factors by both analytic and approximation methods. Nevertheless, this leads to a large number of combinations and functions.
- A single function masize was created in R/gap to simplify this.

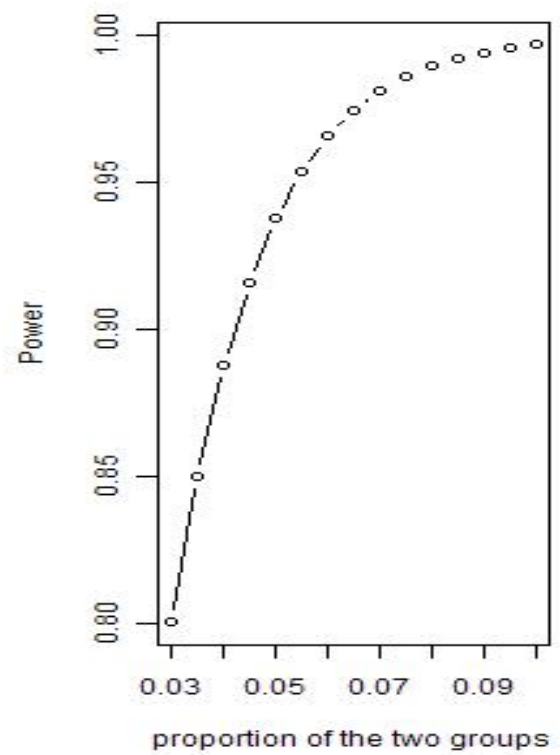
Two-stage design on main effect

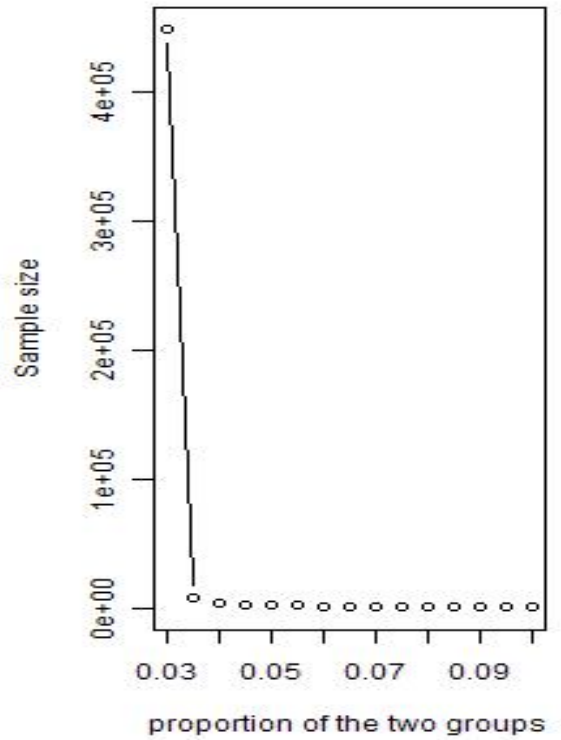
- The goal is to reduce cost without compromising efficiency. Given our study sample and SNPs of interest are defined, a staged design furnishes collection of all information in several steps.
- In the simplest and well-studied two-staged design of genetic case-controls studies, a proportion of individuals is genotyped at all of the SNPs and a proportion of the most significant ones is selected and to be carried over as replication study at the second stage. Skol et al. Nat Genet 2006; 38(2):209-13 (check the associate website for a program called CaTS).
- It was implemented in the function tscc within R/gap.

Some complications-two-stage GEI

- A case-only design is used as the first stage.
- This is to be followed by a second stage involving both cases and controls.
- Some recent references are given here:
  - Kass PH, Gold EB. in Ahrens W, Pigeot I (Eds) Handbook of Epidemiology 2005; I.7
  - Murcray CE, et al. Am J Epidemiol 2008; 169:219-26
  - Li D, Conti DV. Am J Epidemiol 2008; 169:497-504
  - Kooperberg C, LeBlanc M. Genet Epidemiol 2008; 32:255-63
  - Thomas DC, et al. Stat Sci 2009; 24:414-29
  - Schaid DJ, Sinnwell JP. Hum Genet 2010; 127:659-68

A flaw in the case-cohort framework for extremely large cohort?





The right panels show when the cohort of 6.5million, the power/sample size is unstable such that a change of p1 from 0.04 to 0.03 led to sample size increase from 3968 to 211,480!

Moreover

- The formula is only appropriate for the case of dominant model, and it would be much preferable to consider the most widely used additive model.
- We will need probability weighting to allow for general genetic models to be specified. This has not been established but would be analogous to the framework as implemented in the computer program Quanto which is widely used.

Other packages

- powerGWASinteraction (Kooperberg & LeBlanc 2008). It calculates power for SNP x SNP and SNP x environment interactions in genome-wide association studies. It assumes a two-stage analysis, where only SNPs that are significant at a marginal significance level α_1 are investigated for interactions, and a binary environmental covariate.
- trex (Schaid & Sinnwell 2010). It implements truncated exact test for two-stage case-control design for studying rare genetic variants. It consists of a screening stage focusing on rare variants in cases by which number of case-carriers of any rare variants exceeds a user-speficified threshold will have additional cases and controls genotyped and analysed for all cases and controls in the second stage.

Summary

- Power calculation is an integrated part of in designing epidemiological studies and closely linked with the statistical analysis to be carried out.
- There are certain basic principles to follow whenever new problem comes along. We can then implement in R.
- Customised software, however, would greatly facilitate the calculation, e.g.,
  - R - for the comparison between association/linkage including the Mendelian randomisation example.
  - SAS - for the regression example.
  - Quanto - for GEI here.

Return to the beginning

- The statistical understanding is on the evolving, e.g., author of Quanto is currently conducting a survey, which should lead to amendment to our calculation for gene-environment interaction here.
- The mult-stage approach includes case-only coupled with case-control samples and current focus on rare variants.
- There appears to have problems with case-cohort sampling from extremely large cohort.
- The case of Mendelian randomisation should be considered in a broader framework, e.g., survival outcome.

References for Genetic Epidemiology

- Armitage P, Colton T. Encyclopedia of Biostatistics, Second Edition, Wiley 2005
- Balding DJ, Bishop M, Cannings C. Handbook of Statistical Genetics, Third Edition, Wiley 2007
- Elston RC, Johnson W. Basic Biostatistics for Geneticists and Epidemiologists: A Practical Approach. Wiley 2008
- Haines JL, Pericak-Vance M. Genetic Analysis of Complex Diseases, Second Edition. Wiley 2006
- Rao DC, Gu CC (Eds). Genetic Dissection of Complex Traits, Second Edition. Academic Press 2008
- Thomas DC. Statistical Methods for Genetic Epidemiology. Oxford University Press 2004
- Speicher M, Antonarakis SE, Motulsky AG. Vogel and Motulskys Human Genetics-Problems and Approaches, Springer 2009.

References for Mendelian Randomisation

- Mackinnon DP. Introduction to Statistical Mediation Analysis. Taylor & Francis Group, LLC 2008.
- Palmer TM, et al. Adjusting for bias and unmeasured confounding in Mendelian randomization studies with binary responses. Int J Epidemiol 2008;37:1161-68
- Vittinghoff E, et al. Sample size calculations for evaluating mediation. Stat Med 2009; 28:541-57
- Mckeigue PM, et al. Bayesian methods for instrumental variable analysis with genetic instruments (Mendelian randomization): example with urate transporter *SLC2A9* as an instrumental variable for effect of urate levels on metabolic syndrome. Int J Epidemiol 2010; 39:907-18

Acknowledgements

The results presented here/hereafter were based on work recently done at MRC.

- Functions pbsize and fbsize were originally written in C at IoP and refined before and after the JSS paper.
- Function ccsize for case-cohort design attributes to EPIC-Norfolk obesity study, and function tscc for stage design was implemented as a result of an internal journal club.
- The Mendelian randomisation example was used in preparation for MRC QQR.
- R^2 regression methods attributes to ELSA-DNAR applications from UCL (http://www.natcen.ac.uk/elsa/).
- Case-control results were obtained for InterAct (http://www.inter-act.eu/).

Power estimation based on proportion of variance explained

	R^2				
Sample size	0.1	0.2	0.3	0.4	0.5
$\alpha = 10^{-5}$					
10,000	0.10	0.52	0.86	0.97	1
15,000	0.29	0.86	0.99	1.00	1
20,000	0.52	0.97	1.00	1.00	1
25,000	0.72	1.00	1.00	1.00	1
$\alpha = 5 \times 10^{-7}$					
10,000	0.031	0.29	0.67	0.9	0.98
15,000	0.124	0.67	0.95	1.0	1.00
20,000	0.290	0.90	1.00	1.0	1.00
25,000	0.489	0.98	1.00	1.0	1.00

SAS code


```
proc power;
ods output output=op;
multreg model = fixed
        alpha = 0.00001 0.000001 0.0000005
        nfullpred = 1
        ntestpred = 1
        rsqfull = 0.001 to 0.005 by 0.001
        rsqdiff = 0.001 to 0.005 by 0.001
        ntotal = 10000 to 25000 by 1000 power = .;
run;
```

Note that ods can suppress/save all outputs to databases.

Models for gene-environment interaction

Let D=disease, E=exposure, and g=genotype at a candidate locus with susceptibility allele A and normal allele a. The population prevalence of A and exposure will be denoted by q_A and p_E. Under HWE, the genotypes AA, Aa, aa have frequencies $P(g|q_A)$=q_A^2, $2q_A(1-q_A)$, $(1-q_A)^2$. We assume particular model e.g., $G(g) = 0, 1, 2$ and relate disease to genetic and environmental covariates through logistic and log-linear models

$$P(D = 1|G, E) = \frac{e^{\alpha+\beta_g G+\beta_e E+\beta_{ge} GE}}{1 + e^{\alpha+\beta_g G+\beta_e E+\beta_{ge} GE}}$$

and

$$P(D = 1|G, E) = e^{\alpha+\beta_g G+\beta_e E+\beta_{ge} GE}$$

so that the baseline probabilities of disease in the population is given by $e^{\alpha}/(1 + e^{\alpha})$ and e^{α} whereas e_g^{β}, e_e^{β}, e_{ge}^{β} are the genetic, environmental and interactive relative risks.

Power calculation under matched design

- We can use conditional logistic regression model

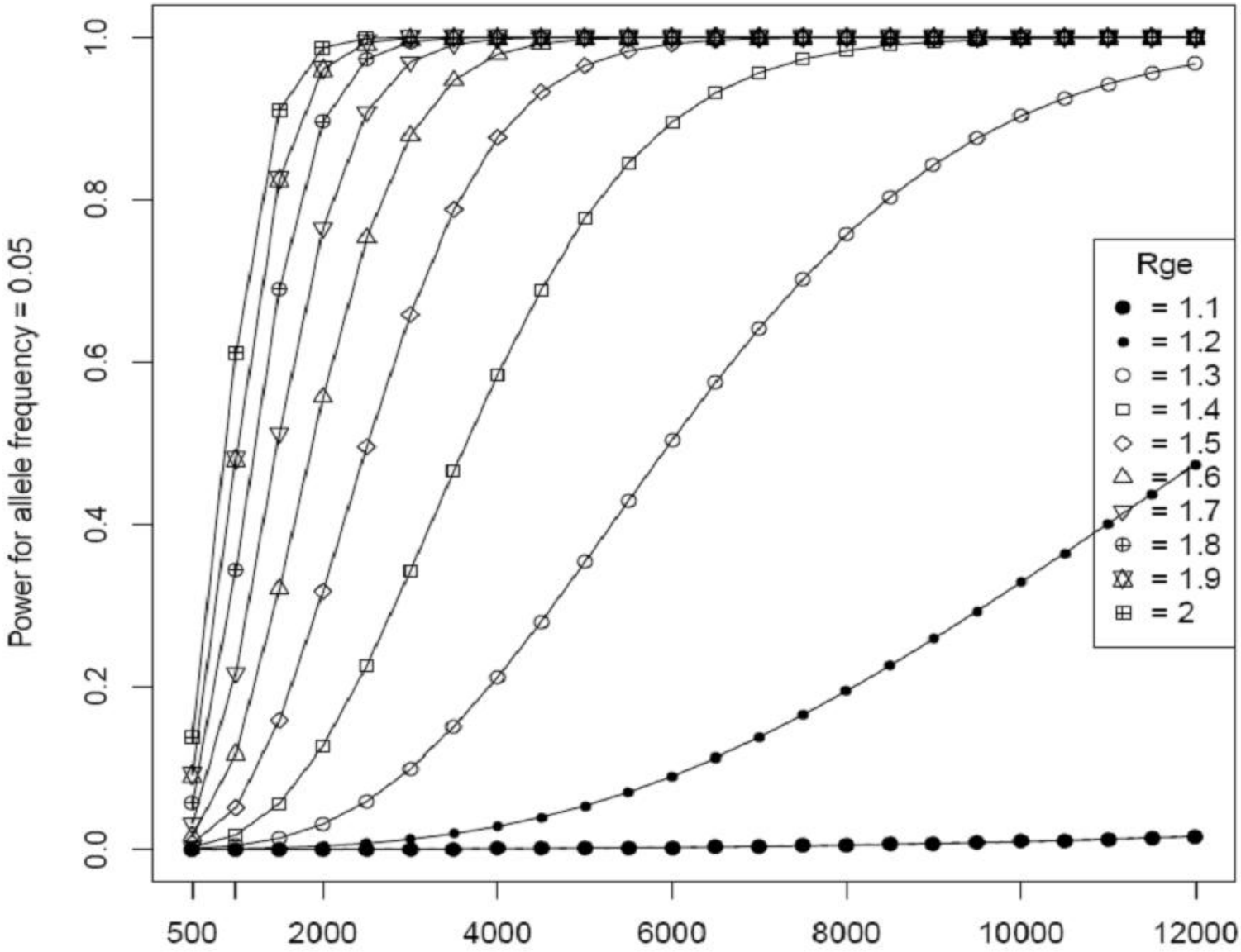
$$L(\beta_g, \beta_e, \beta_{ge}) = \prod_{i=1}^{N} \frac{e^{\beta_g G_{ij} + \beta_e E_{ij} + \beta_{ge} G_{ij} E_{ij}}}{\sum_{j \in M(i)} e^{\beta_g G_{ij} + \beta_e E_{ij} + \beta_{ge} G_{ij} E_{ij}}}$$

  where $M(i)$ includes all subjects in matched set i.
- Power/sample size calculation can proceed with contrasting $l^1 = \ln[L(\beta_g, \beta_e, \beta_{ge})]$, $l^0 = \ln[L(\beta_g, \beta_e)]$ with $\Lambda = 2(\hat{l}^1 - \hat{l}^0)$ and $N\Lambda$ being the non-centrality parameter of chi-squared distribtion under the alternative hypothesis.
- The required sample size is obtained via equating the noncentrality parameter to theoretical values under a given significant level and power (the previous conceptual picture still applies).

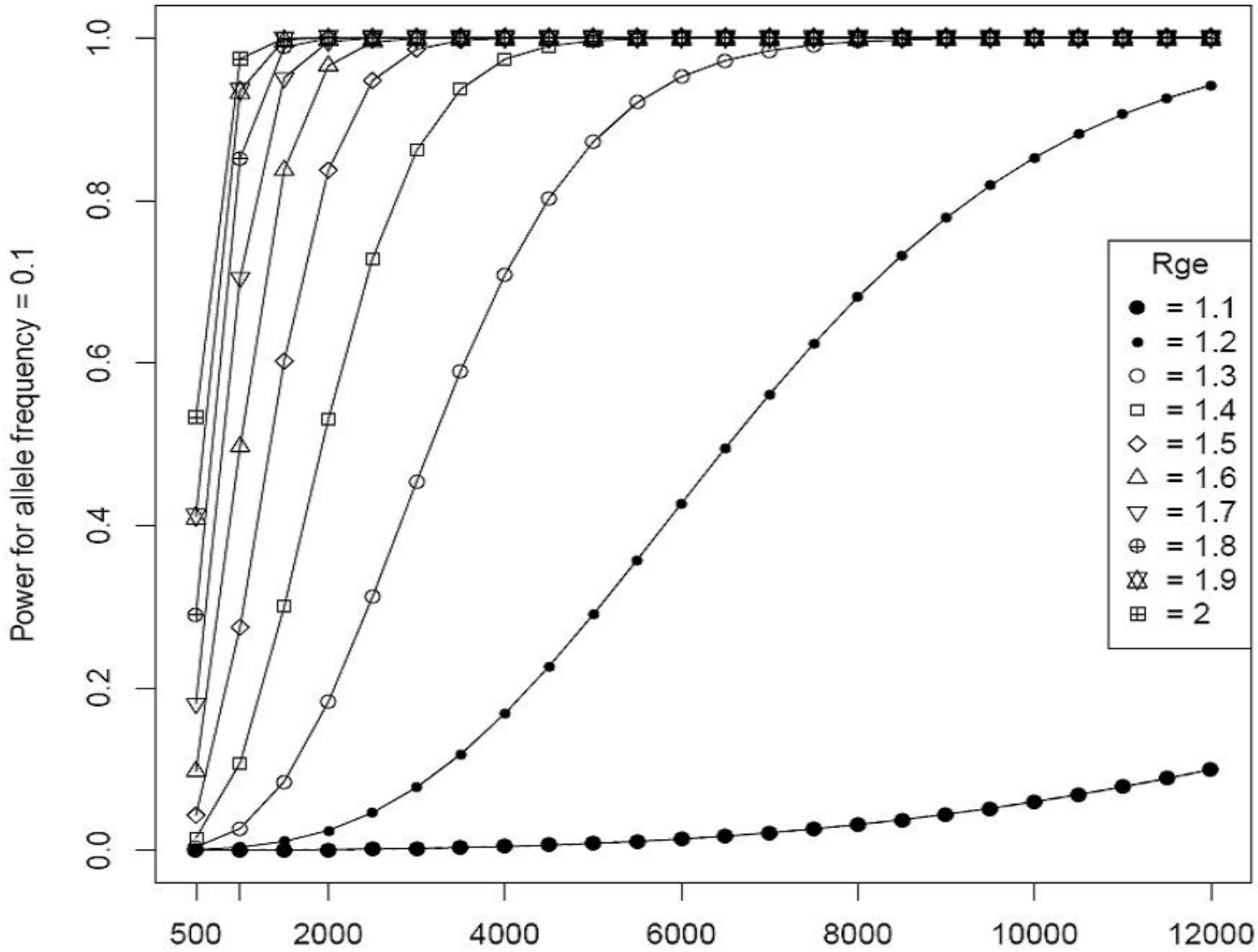
Gene-environment interaction in T2D

- Legends in the project manual were perhaps confusing so it is worthwhile to re-present here.
  - Matched case-control study
  - Type I error rate (α) = 0.00001 (two-sided)
  - Continuous environmental factors with standard deviation 1, and interaction odds ratio (R_{ge}) = 1.2 - 4
  - K = 0.05 (0.1 - 0.15)
  - Sample size (N) = 500 - 12,000
  - Additive model
  - Allele frequency (p) = 0.05, 0.1, 0.2, 0.3
- We supplied these to Quanto 1.0 (http://hydra.usc.edu/gxe, now available on the Epidemiology Unit machines) Gauderman WJ. Stat Med 21:35-50, 2002

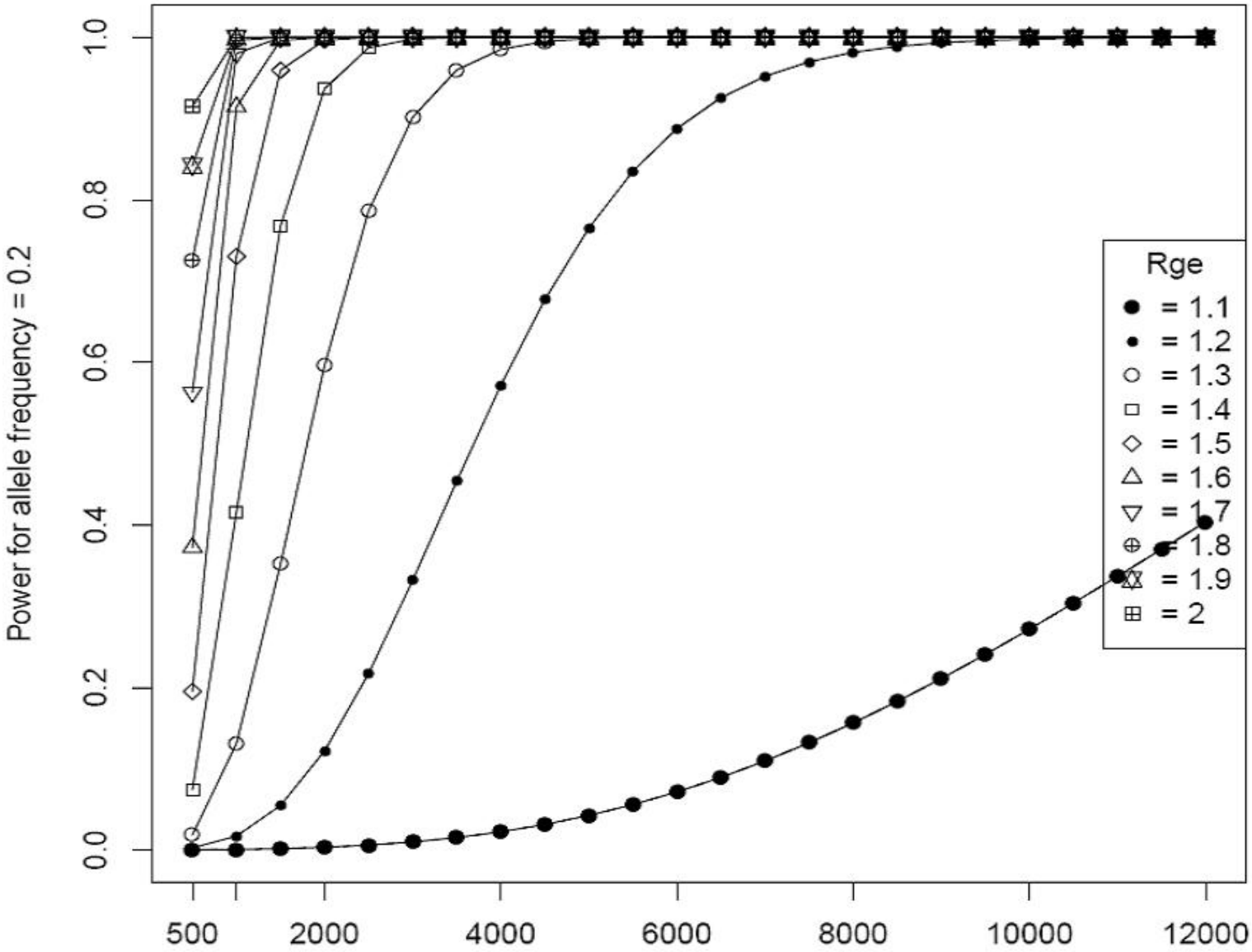
Scenario 1



Scenario 2



Scenario 3



Scenario 4