Method to estimate the approximate samples size that yield a certain number of significant GWAS signals in polygenic traits

Silviu-Alin Bacanu[1*] and Kenneth S. Kendler[1]
[1]Psychiatric Department, Virginia Commonwealth University

Abstract

To argue for increased sample collection for disorders without significant findings, researchers retorted to plotting, for multiple traits, the number of significant findings as a function of the sample size. However, for polygenic traits, the prevalence of the disorder confounds the relationship between the number of significant findings and the sample size. To adjust the number of significant findings for prevalence, we develop a method that uses the expected noncentrality of the contrast between liabilities of cases and controls. We empirically find that, when compared to the sample size, this measure is a better predictor of number of significant findings. Even more, we show that the sample size effect on the number of signals is explained by the noncetrality measure. Finally, we provide an R script to estimate the required sample size (non-centrality) needed to yield a pre-specified number of significant findings.

*Correspondence: silviu-alin.bacanu@vcuhealth.org

To illustrate the tractability of complex diseases, researchers intuitively plot/regress[1,2] the number of significant findings, n_s, by the sample size, N, (see Fig.2 in Kim et al[1] and Fig. 3 in this paper). Early in the GWAS era such a plot suggested that the number of significant hits is approximately linear after the emergence of the first genome wide significant finding (Mark Daly PGC presentation). While such analyses are definitely informative, for polygenic traits such plots are confounded by the trait prevalences (Fig. 2 in Gratten et all[3]). For a better characterization of trait effect size that is not cryptically influenced by prevalence, we propose an approach to adjust traits for their prevalences and provide an empirical relation between such normalized variables and the number of significant findings for given sample sizes.

Let us assume the existence of biologically informative covariates, e.g. gender and ancestry principal components, which helps us in recovering the liability to disease (even up to a multiplication factor), L, for both cases and controls for a binary trait (BT) of prevalence K. (It should be noted that working on the liability scale, instead of the natural binary case control scale, is also supported by the Invariance Principle of statistical mathematics[4], which states that the inference should not be affected by the scale/transformations one chooses to employ.) For the threshold-liability model, let the threshold be $\tau_K = \Phi^{-1}(1-K)$, where Φ^{-1} is the inverse cumulative distribution function of the Gaussian distribution. In a threshold-liability model $L \geq \tau_K$ for cases and $L < \tau_K$ for controls. Thus, for a study consisting of N_1 cases and N_2 controls, the normalized effect size (δ), i.e. the difference in liability between cases and controls after adjusting for its standard error, is:

$$\delta(N_1, N_2) = \frac{E(L|L\geq\tau_K) - E(L|L<\tau_K)}{\sqrt{\frac{Var(L|L\geq\tau_K)}{N_1} + \frac{Var(L|L<\tau_K)}{N_2}}} \quad (1).$$

If φ is the probability density function for the Gaussian distribution, after substituting the expressions for expectation and variance of truncated Gaussian distributions[5], relationship (1) becomes:

$$\delta(N_1, N_2) = \frac{\frac{\varphi(\tau_K)}{1-\Phi(\tau_K)} - \frac{-\varphi(\tau_K)}{\Phi(\tau_K)}}{\sqrt{\frac{1+\tau_K\frac{\varphi(\tau_K)}{1-\Phi(\tau_K)} - \left(\frac{\varphi(\tau_K)}{1-\Phi(\tau_K)}\right)^2}{N_1} + \frac{1-\tau_K\frac{\varphi(\tau_K)}{\Phi(\tau_K)} - \left(\frac{\varphi(\tau_K)}{\Phi(\tau_K)}\right)^2}{N_2}}} =$$

$$\frac{\frac{\varphi(\tau_K)}{K} + \frac{\varphi(\tau_K)}{1-K}}{\sqrt{\frac{1+\tau_K\frac{\varphi(\tau_K)}{K} - \left(\frac{\varphi(\tau_K)}{K}\right)^2}{N_1} + \frac{1-\tau_K\frac{\varphi(\tau_K)}{1-K} - \left(\frac{\varphi(\tau_K)}{1-K}\right)^2}{N_2}}} \quad (2).$$

However, most often researchers work with the χ^2 distribution and, on this scale, the non-centrality parameter of contrasting case and control liabilities is $\lambda(N_1, N_2) = \delta^2(N_1, N_2)$. In turn, detection power is increasing with increased non-centrality parameter.

While equation (2) is derived for binary traits, it can be extended to quantitative traits (QT). For instance, we can use a first order approximation for QT as a case control trait with prevalence of 50% (i.e. a contrast above median height vs. below median height). While, in practice, such a discretization approach leads to power loss, we stress that the GWAS statistics are already obtained using a QT. The above/below median approximation is only used to extend the use of equation (2). With this preparatory work, the noncentrality per case and control unit ($N_1 = N_2 = 1$), $\lambda(1,1)$, increases by ~60% with a decrease in prevalence (Fig. 1).

To empirically investigate whether λ is a better measure than N_1, or $N = N_1 + N_2$, to describe observed n_s, we analyze the number of significant findings (Table 1) for multiple studies for some of the most widely investigated traits. Three phenotypes are chosen from each of the four investigated trait classes (see Table 1 for references): anthropometric (all QTs) and psychiatric, neurodegenerative and immune diseases (all BT). Anthropometric traits (denoted as Anthro in plot legends) are height (H), body mass index (BMI) and waist-to-hip ratio (WHR). Psychiatric (Psych) traits are the main psychiatric disorders: schizophrenia (SCZ), bipolar disorder (BD) and major depressive disorder (MDD). As neurodegenerative (Neuro) we chose Alzheimer's disease (ALZ), Parkinson's disease (PD and multiple sclerosis (MS). Finally, we chose as immune (Immune) disorders: Crohn's disease (CD), rheumatoid arthritis (RA) and type 2 diabetes (T2D).

To assist in predicting n_s as a function of λ, we also need to determine what transformation should we use for n_s and λ/N_1 to make the relationship between n_s and λ stronger. As mentioned in the introduction, the intuitive idea is to use the identity scale, i.e. no transformation. However, given that n_s can be viewed as a sum of Bernoulli variables (0- non-significant and 1 significant) , Chernoff inequality[4] suggests that a log transformation of n_s is likely much more desirable. For effect sizes λ (and likely, as its transformation, N), the plotting of the log probability of a significant signal ($\alpha = 5x10^{-8}$) as a function of noncentrality, λ, and its log transformation, also show a much better fit (Fig. 2) for the log transformation (R^2 of 99.4% vs 91%). Given that the probability of a significant find is proportional to the number of significant findings, this suggests that the log transformation is also suitable for λ.

Thus to establish the relationship between regressing $\log[n_s]$ and $\log(\lambda)$ (also $\log N_1, \log N$) we use a gls model (in nlme R package) assuming an autoregressive of order 1 (AR1) correlation structure for observations within the same trait (due to earlier studies being included in all subsequent meta-analysis of this disease). We used the model to test whether the effects of N and N_1 on n_s are mediated only via $\log(\lambda)$, i.e. we regressed $\log[n_s]$ on $\log(\lambda)$, $\log[N]$, $\log(N_1)$, N and N_1. In this model, only $\log(\lambda)$ was significant (p-value of 0.025) and all the others were not (p-values of 0.58 and 0.73). Even more, stepwise elimination on non-significant variables left only $\log(\lambda)$ as significant with $\log[N]$ being the last to be eliminated with a p-values of 0.65. This result strongly suggests that the effect of N and N_1 on n_s is wholly mediated by λ and thus non-centrality is a better predictor than sample size. The gls model was also used to vividly illustrate the better performance of our theoretically chosen transformations: when using the natural sample size scale for both n_s and λ (Fig. 3), the fit (R^2=0.42) is much poorer than using log scale for both (Fig. 4) (R^2=0.71). (The similar in spirit square root transformation of λ performed only moderately worse than log.)

We stress again that the above results suggest that our proposed measure on log scale better predicts the (log) number of significant findings for traits of various prevalences. Thus, λ from relationship (1) is a desirable effect size measure that is not confounded by prevalence. Based on the gls regression of $\log[n_s]$ on $\log(\lambda)$, the best prediction for the number of significant findings is:

$$n_s = 5.6\, x\, 10^{-4}\, \lambda^{0.89} = 5.6\, x\, 10^{-4} \left[\frac{\left(\frac{\varphi(\tau_K)}{K} + \frac{\varphi(\tau_K)}{1-K}\right)^2}{\frac{1+\tau_K \frac{\varphi(\tau_K)}{K}\left(\frac{\varphi(\tau_K)}{K}\right)^2}{N_1} + \frac{1-\tau_K \frac{\varphi(\tau_K)}{1-K}\left(\frac{\varphi(\tau_K)}{1-K}\right)^2}{N_2}} \right]^{0.89} \quad (3).$$

However, most of the time the researchers want to estimate the number of cases, N_1, needed to obtain a certain number of significant findings, n_s. To this end let $N_2 = q\ N_1$, where generally $q > 1$ is largely known. Then equality (3) can be solved for N_1, as follows:

$$N_1 = \left(\frac{n_s}{5.6\ x\ 10^{-4}}\right)^{1.124} \frac{\left[1+\tau_K\frac{\varphi(\tau_K)}{K}-\left(\frac{\varphi(\tau_K)}{K}\right)^2+\frac{1-\tau_K\frac{\varphi(\tau_K)}{1-K}-\left(\frac{\varphi(\tau_K)}{1-K}\right)^2}{q}\right]}{\left(\frac{\varphi(\tau_K)}{K}+\frac{\varphi(\tau_K)}{1-K}\right)^2},$$

or $N_1 = 4{,}519\ n_s^{1.124} \frac{\left[1+\tau_K\frac{\varphi(\tau_K)}{K}-\left(\frac{\varphi(\tau_K)}{K}\right)^2+\frac{1-\tau_K\frac{\varphi(\tau_K)}{1-K}-\left(\frac{\varphi(\tau_K)}{1-K}\right)^2}{q}\right]}{\left(\frac{\varphi(\tau_K)}{K}+\frac{\varphi(\tau_K)}{1-K}\right)^2}$ (4). To assist applied researchers in sizing their studies, we present in the Appendix the R implementation of equalities (3) and (4).

Appendix

 # R function for estimating the noncentrality and # cases

K- prevalence, Nca - # cases ; Nco - # controls

```
get.nonc<-function(K=K, Nca=Nca, Nco=Nco){
    tau.K<-qnorm(K, low=F)
    ncp<-(dnorm(tau.K)*(1/K+1/(1-K)))^2/((1+tau.K*dnorm(tau.K)/K-
(dnorm(tau.K)/K)^2)/Nca+(1-tau.K*dnorm(tau.K)/(1-K)-(dnorm(tau.K)/(1-K))^2)/Nco)
    ncp
}
```

 # R function for estimating the required # cases yielding # of signals using our formula (4)

K- prevalence, ns - # desired significant findings & q=ratio of controls to cases (often q>1)

```
get.n.cases<-function(K=K, ns=1, q=1){
    tau.K<-qnorm(K, low=F)
    N1<-4519*ns^1.124*((1+tau.K*dnorm(tau.K)/K-(dnorm(tau.K)/K)^2)+(1-
tau.K*dnorm(tau.K)/(1-K)-(dnorm(tau.K)/(1-K))^2)/q)/(dnorm(tau.K)*(1/K+1/(1-K)))^2
    N1
}
```

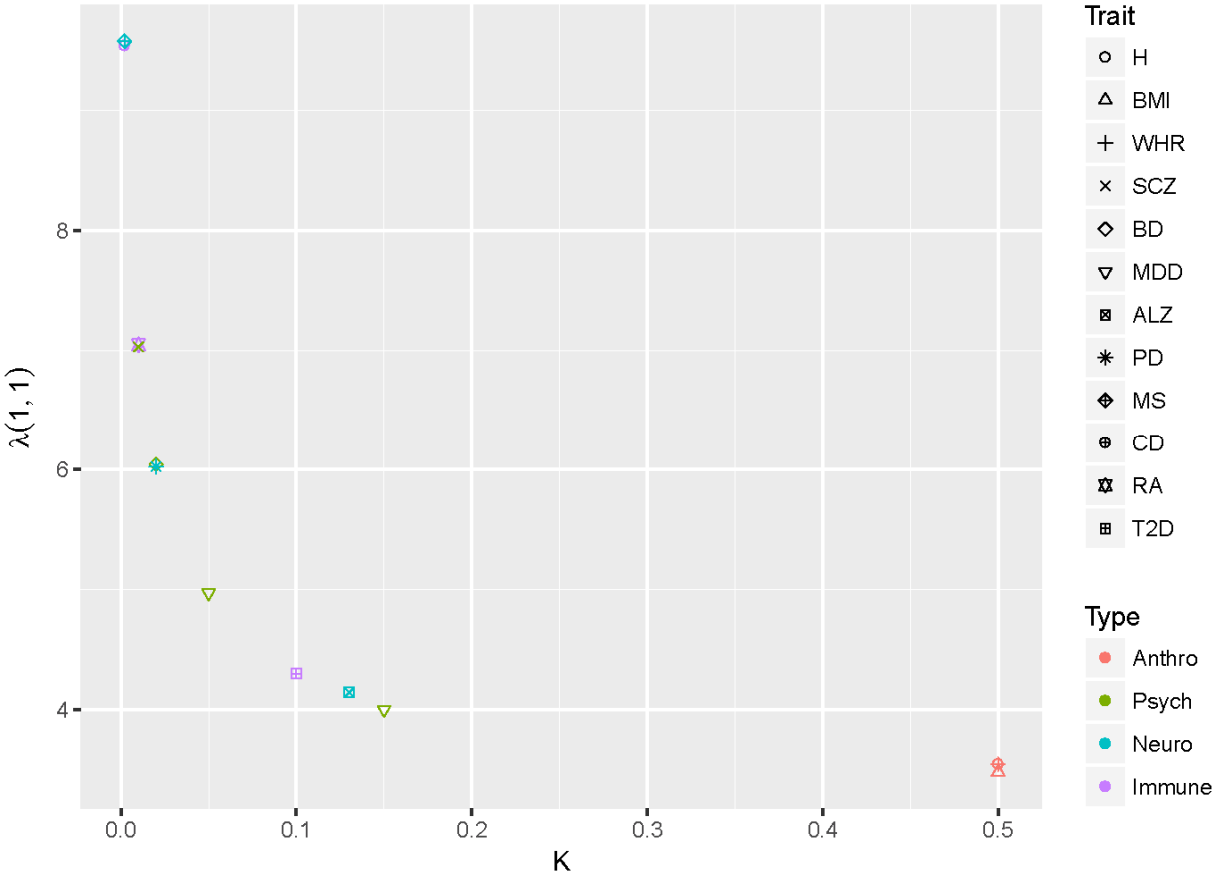

Figure 1. Noncentrality parameter for various traits

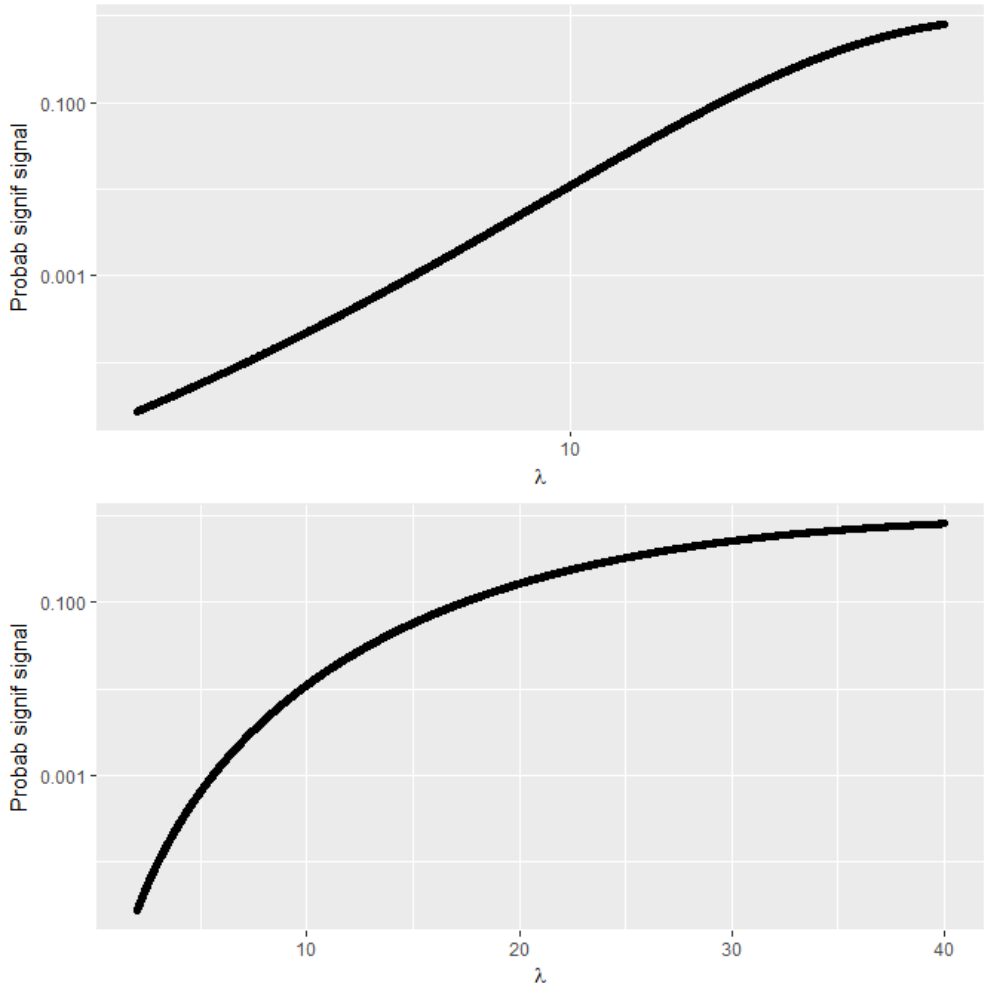

Figure 2. The probability of a significant signal (log scale) as a function of noncentrality on log scale (above) and identity scale (below).

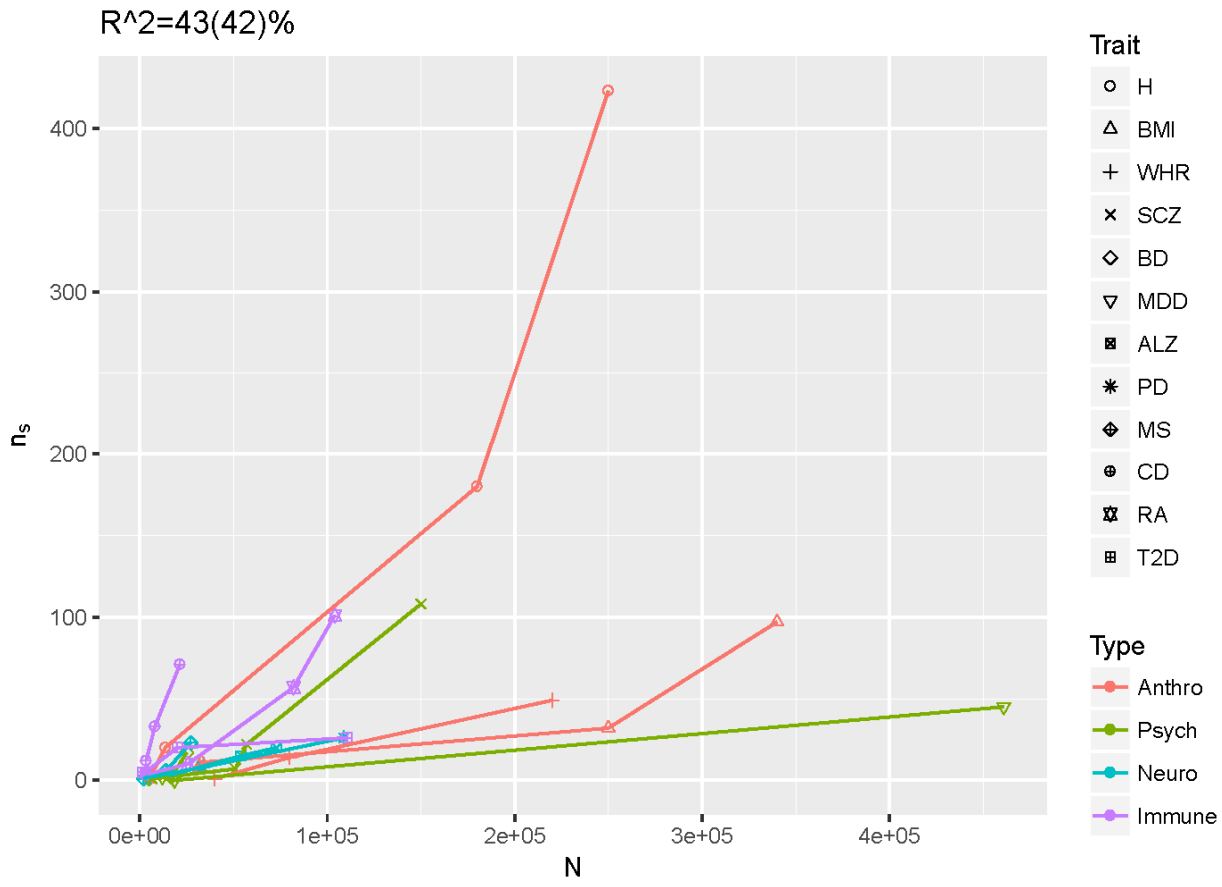

Figure 3. Number of significant findings vs. sample size (without type 2 diabetes-T2D).

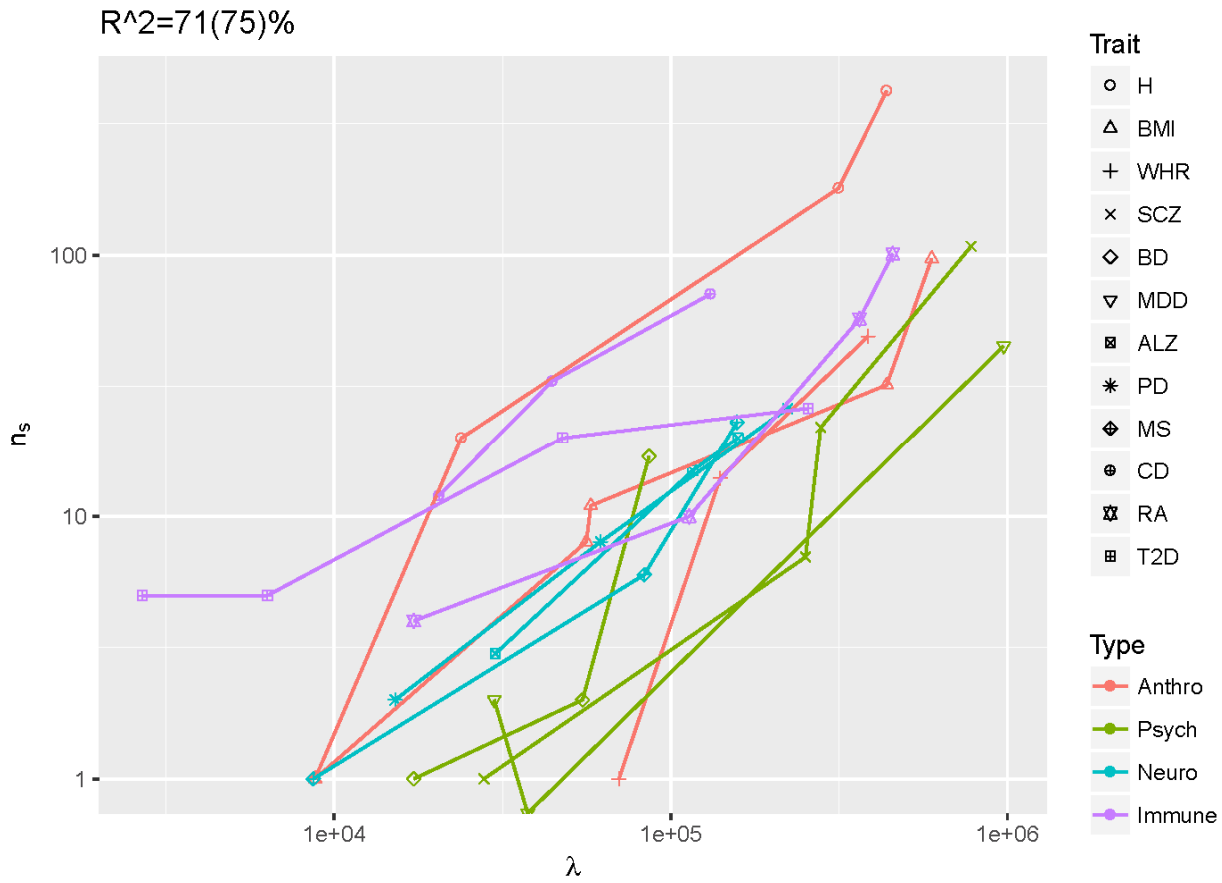

Figure 4. Number of significant findings vs noncentrality parameter (without T2D). Both axes are log scale.

Table 1. Table of Studies

Trait	Abbrev.	K	N cases	N controls	Ns	First author and reference
			6,800	6,800	20	Weedon[6]
Height	H	0.5	90,000	90,000	180	Lango[7]
			125,000	125,000	423	Wood[8]
			2,500	2,500	1	Scuteri[9]
			16,000	16,000	8	Willer[10]
Body Mass Index	BMI	0.5	16,500	16,500	11	Thorleifsson[11]
			125,000	125,000	32	Speliotes[12]
			170,000	170,000	97	Locke[13]
			20,000	20,000	1	Lindgren[14]
Waist to Hip Ratio	WHR	0.5	40,000	40,000	14	Heid[15]
			110,000	110,000	49	Shungin[16]
			17,500	33,500	7	PGC1[17]
Schizophrenia	SCZ	0.01[18]	20,000	37,000	22	PGC1.5[19]
			37,000	113,000	108	PGC2[20]
			2,000	3,000	1	WTCCC[21]
Bipolar Disorder	BD	0.02[22]	7,500	9,250	2	PGC1[23]
			10,000	15,000	5	Muhleisen (personal communication)
Major Depressive Disorder	MDD	0.15[24]	9,000	9,500	0	PGC MDD[25]
(Recurrent MDD)		(0.05)[26]	6,000	6,000	2	CONVERGE[26]
			131,000	330,000	45	PGC2 MDD (online presentation)
			8,300	7,300	3	Naj[27]
Alzheimer's Disease	ALZ	0.13[27]	17,000	37,000	15	Lambert[28]
			25,000	48,000	20	Lambert[28]
			1,700	4,000	2	Simon-Sanchez[29]
Parkinson's Disease	PD	0.02[30]	3,500	30,000	8	Do[31]
			13,700	95,000	26	Nalls[30]
			1,000	900	1	Baranzini[32]
Multiple Sclerosis	MS	0.002[32]	4,800	9,300	6	De Jager[33]
			9,800	17,400	23	IMSGC[34]
			1,000	2,345	12	McGovern[35]
Crohn's Disease	CD[35]	0.002	3,250	4,800	33	Barrett[36]
			6,350	15,050	71	Franke[37]
			2,100	2,500	4	Jiang[38]
Rheumatoid Arthritis	RA	0.01[38]	5,500	20200	10	Stahl[39]
			20,000	620,00	57	Okada[40]
			661	614	5	Sladek[41]
Type 2 Diabetes	T2D	0.1[41]	1,464	1,467	5	Diabetes at BROAD[42]
			5,500	14,500	20	Kooner[43]
			26,500	84,000	26	DIAGRAM[44]

1. Kim, Y., Zerwas, S., Trace, S.E. & Sullivan, P.F. Schizophrenia genetics: where next? *Schizophr Bull* **37**, 456-63 (2011).
2. Visscher, P.M., Brown, M.A., McCarthy, M.I. & Yang, J. Five years of GWAS discovery. *Am. J Hum. Genet* **90**, 7-24 (2012).
3. Gratten, J., Wray, N.R., Keller, M.C. & Visscher, P.M. Large-scale genomics unveils the genetic architecture of psychiatric disorders. *Nat Neurosci* **17**, 782-90 (2014).
4. Casella, G. & Berger, R.L. *Statistical Inference*, (Brooks/Cole Publishing Company, 1990).
5. Johnson, N.L., Kotz, S. & Balakrishnan, N. *Continuous univariate distributions*, (Wiley, New York, 1994).
6. Weedon, M.N. *et al.* Genome-wide association analysis identifies 20 loci that influence adult height. *Nat Genet* **40**, 575-83 (2008).
7. Lango, A.H. *et al.* Hundreds of variants clustered in genomic loci and biological pathways affect human height. *Nature* **467**, 832-838 (2010).
8. Wood, A.R. *et al.* Defining the role of common variation in the genomic and biological architecture of adult human height. *Nat Genet* **46**, 1173-86 (2014).
9. Scuteri, A. *et al.* Genome-wide association scan shows genetic variants in the FTO gene are associated with obesity-related traits. *PLoS Genet* **3**, e115 (2007).
10. Willer, C.J. *et al.* Six new loci associated with body mass index highlight a neuronal influence on body weight regulation. *Nat. Genet* **41**, 25-34 (2009).
11. Thorleifsson, G. *et al.* Genome-wide association yields new sequence variants at seven loci that associate with measures of obesity. *Nat Genet* **41**, 18-24 (2009).
12. Speliotes, E.K. *et al.* Association analyses of 249,796 individuals reveal 18 new loci associated with body mass index. *Nat. Genet* **42**, 937-948 (2010).
13. Locke, A.E. *et al.* Genetic studies of body mass index yield new insights for obesity biology. *Nature* **518**, 197-206 (2015).
14. Lindgren, C.M. *et al.* Genome-wide association scan meta-analysis identifies three Loci influencing adiposity and fat distribution. *PLoS Genet* **5**, e1000508 (2009).
15. Heid, I.M. *et al.* Meta-analysis identifies 13 new loci associated with waist-hip ratio and reveals sexual dimorphism in the genetic basis of fat distribution. *Nat Genet* **42**, 949-60 (2010).
16. Shungin, D. *et al.* New genetic loci link adipose and insulin biology to body fat distribution. *Nature* **518**, 187-96 (2015).
17. Ripke, S. *et al.* Genome-wide association analysis identifies 13 new risk loci for schizophrenia. *Nat Genet* **45**, 1150-1159 (2013).
18. Saha, S., Chant, D., Welham, J. & McGrath, J. A systematic review of the prevalence of schizophrenia. *PLoS Med* **2**, e141 (2005).
19. Ripke, S. *et al.* Genome-wide association study identifies five new schizophrenia loci. *Nat. Genet* **43**, 969-976 (2011).
20. Consortium, S.W.G.o.t.P.G. Biological insights from 108 schizophrenia-associated genetic loci. *Nature* **511**, 421-427 (2014).
21. Consortium, W.T.C.C. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* **447**, 661-678 (2007).

22. Grant, B.F. *et al.* Prevalence, correlates, and comorbidity of bipolar I disorder and axis I and II disorders: results from the National Epidemiologic Survey on Alcohol and Related Conditions. *J Clin Psychiatry* **66**, 1205-15 (2005).
23. Sklar, P. *et al.* Large-scale genome-wide association analysis of bipolar disorder identifies a new susceptibility locus near ODZ4. *Nat. Genet* **43**, 977-983 (2011).
24. Blazer, D.G., Kessler, R.C., McGonagle, K.A. & Swartz, M.S. The prevalence and distribution of major depression in a national community sample: the National Comorbidity Survey. *Am J Psychiatry* **151**, 979-86 (1994).
25. Major Depressive Disorder Working Group of the Psychiatric, G.C. *et al.* A mega-analysis of genome-wide association studies for major depressive disorder. *Mol Psychiatry* **18**, 497-511 (2013).
26. Consortium, C. Sparse whole-genome sequencing identifies two loci for major depressive disorder. *Nature* **523**, 588-591 (2015).
27. Naj, A.C. *et al.* Common variants at MS4A4/MS4A6E, CD2AP, CD33 and EPHA1 are associated with late-onset Alzheimer's disease. *Nat Genet* **43**, 436-41 (2011).
28. Lambert, J.C. *et al.* Meta-analysis of 74,046 individuals identifies 11 new susceptibility loci for Alzheimer's disease. *Nat Genet* **45**, 1452-8 (2013).
29. Simon-Sanchez, J. *et al.* Genome-wide association study reveals genetic risk underlying Parkinson's disease. *Nat Genet* **41**, 1308-12 (2009).
30. Nalls, M.A. *et al.* Large-scale meta-analysis of genome-wide association data identifies six new risk loci for Parkinson's disease. *Nat Genet* **46**, 989-93 (2014).
31. Do, C.B. *et al.* Web-based genome-wide association study identifies two novel loci and a substantial genetic component for Parkinson's disease. *PLoS Genet* **7**, e1002141 (2011).
32. Baranzini, S.E. *et al.* Genome-wide association analysis of susceptibility and clinical phenotype in multiple sclerosis. *Hum Mol Genet* **18**, 767-78 (2009).
33. De Jager, P.L. *et al.* Meta-analysis of genome scans and replication identify CD6, IRF8 and TNFRSF1A as new multiple sclerosis susceptibility loci. *Nat Genet* **41**, 776-82 (2009).
34. International Multiple Sclerosis Genetics, C. Network-based multiple sclerosis pathway analysis with GWAS data from 15,000 cases and 30,000 controls. *Am J Hum Genet* **92**, 854-65 (2013).
35. McGovern, D.P. *et al.* Fucosyltransferase 2 (FUT2) non-secretor status is associated with Crohn's disease. *Hum Mol Genet* **19**, 3468-76 (2010).
36. Barrett, J.C. *et al.* Genome-wide association defines more than 30 distinct susceptibility loci for Crohn's disease. *Nat Genet* **40**, 955-62 (2008).
37. Franke, A. *et al.* Genome-wide meta-analysis increases to 71 the number of confirmed Crohn's disease susceptibility loci. *Nat Genet* **42**, 1118-25 (2010).
38. Jiang, L. *et al.* Novel risk loci for rheumatoid arthritis in Han Chinese and congruence with risk variants in Europeans. *Arthritis Rheumatol* **66**, 1121-32 (2014).
39. Stahl, E.A. *et al.* Genome-wide association study meta-analysis identifies seven new rheumatoid arthritis risk loci. *Nat Genet* **42**, 508-14 (2010).
40. Okada, Y. *et al.* Genetics of rheumatoid arthritis contributes to biology and drug discovery. *Nature* **506**, 376-81 (2014).
41. Sladek, R. *et al.* A genome-wide association study identifies novel risk loci for type 2 diabetes. *Nature* **445**, 881-885 (2007).

42. Diabetes Genetics Initiative of Broad Institute of, H. *et al.* Genome-wide association analysis identifies loci for type 2 diabetes and triglyceride levels. *Science* **316**, 1331-6 (2007).
43. Kooner, J.S. *et al.* Genome-wide association study in individuals of South Asian ancestry identifies six new type 2 diabetes susceptibility loci. *Nat Genet* **43**, 984-9 (2011).
44. Replication, D.I.G. *et al.* Genome-wide trans-ancestry meta-analysis provides insight into the genetic architecture of type 2 diabetes susceptibility. *Nat Genet* **46**, 234-44 (2014).