Supplementary Information for:

Pooling/Bootstrap-based GWAS (*pb*GWAS) Identifies New Loci Modifying the Age of Onset in *PSEN1* p.Glu208Ala Alzheimer's Disease

Jorge I. Vélez[1,*], Settara C. Chandrasekharappa[2], Eliana Henao[3], Ariel F. Martinez[1], Ursula Harper[2], MaryPat Jones[2] , Benjamin D. Solomon[1], Liliana Lopez[3], Gloria Garcia[3], Daniel Camilo Aguirre-Acevedo[3], Natalia Acosta-Baena[3], Juan Carlos Correa[4], Carlos Mario Lopera-Gómez[4], Mario César Jaramillo-Elorza[4], Dora Rivera[3], Kenneth S. Kosik[5], Nicholas J. Schork[6], James M. Swanson[7,8], Francisco Lopera[3,*], and Mauricio Arcos-Burgos[1,3,9,#]

[1] Medical Genetics Branch, National Human Genome Research Institute, National Institutes of Health, Bethesda, MD, USA.
[2] Genome Technology Branch, National Human Genome Research Institute, National Institutes of Health, Bethesda, MD, USA.
[3] Grupo de Neurociencias de Antioquia, Facultad de Medicina, Universidad de Antioquia, Medellín, Colombia.
[4] Escuela de Estadística, Universidad Nacional de Colombia, Sede Medellín, Medellín, Colombia.
[5] Neuroscience Research Institute, University of California at Santa Barbara, Santa Barbara, CA, USA.
[6] Department of Biostatistics and Bioinformatics, The Scripps Research Institute, La Jolla, CA, USA.
[7] Department of Psychiatry, Florida International University, Miami, FL, USA.
[8] Child Development Center, University of California at Irvine, Irvine, CA, USA.
[9] Translational Genomics Group, Department of Translational Medicine, John Curtin School of Medical Research, ANU College of Medicine, Biology & Environment, The Australian National University, Canberra, ACT, Australia.

*These authors contributed equally to this work.

[#] Correspondence to be directed to:
Mauricio Arcos-Burgos, M.D., Ph.D.
Associate Professor and Group Leader
Translational Genomics Group
ANU College of Medicine, Biology & Environment
John Curtin School of Medical Research,
The Australian National University,
Building 131 Garran Road,
Office 3.091, Tel: +61 2 61259396
Canberra, ACT, 0200, Australia.
e-mail: Mauricio.arcos-burgos@anu.edu.au

SUPPLEMENTARY METHODS

An Alternative Empirical Method to Determine Rates of Type I Error Probability. Suppose that m independent tests of the same type are applied e.g. allelic frequencies for *m* SNPS are compared between cases and controls; denote P_i the *P*-value for the *i*-th hypothesis, $i=1,2,\ldots,m$. Under the common null hypothesis, $P_1, P_2,\ldots,P_m$ is a random sample of size *m* following a $U(0,1)$ distribution. Let *V* be a random variable with cumulative distribution function (*cdf*) *F*, and $V_{(m)} = \max\{V_1, V_2,\ldots,V_m\}$ be the maximum in a random sample of size *m*. The exact distribution of $V_{(m)}$ is given by:

$$P(V_{(m)} < t) = \{F(t)\}^m \qquad (1)$$

(Casella & Berger, 2001), Note that if *F* is unknown, calculation of (1) is impossible. Using asymptotic theory, we used the alternative method described by Serfling (1990, pp. 89): as $m \to \infty$ (e.g., several hundred thousands of tests are performed), seems intuitive to derive an empirical test to evaluate significant *P*-values for a fixed type I error probability α.Thus far, consider the random variable:

$$D_m = (V_{(m)} - a_m)/b_m \qquad (2)$$

for some constants: $\{a_m\}$ and $\{b_m\}$, the limiting distribution of (2) has one of three forms (Serfling, 1980; pp. 89). Since $V_1=-\log(P_1)$, $V_2=-\log(P_2),\ldots$, $V_m=-\log(P_m) \sim$ Exponential (1) (Devroye, 1986), by choosing $a_m= \log(m)$ and $b_m=1$ it follows that

$$P(V_{(m)} - \log(m) < t) \to \exp\{-\exp(-t)\}, \qquad -\infty < t < \infty, \quad m \to \infty \qquad (3)$$

(Serfling, 1980; pp. 90). Now, let t_c be the critical value, e.g.

$$P(V_{(m)} - \log(m) < t_c) = \alpha \qquad (4)$$

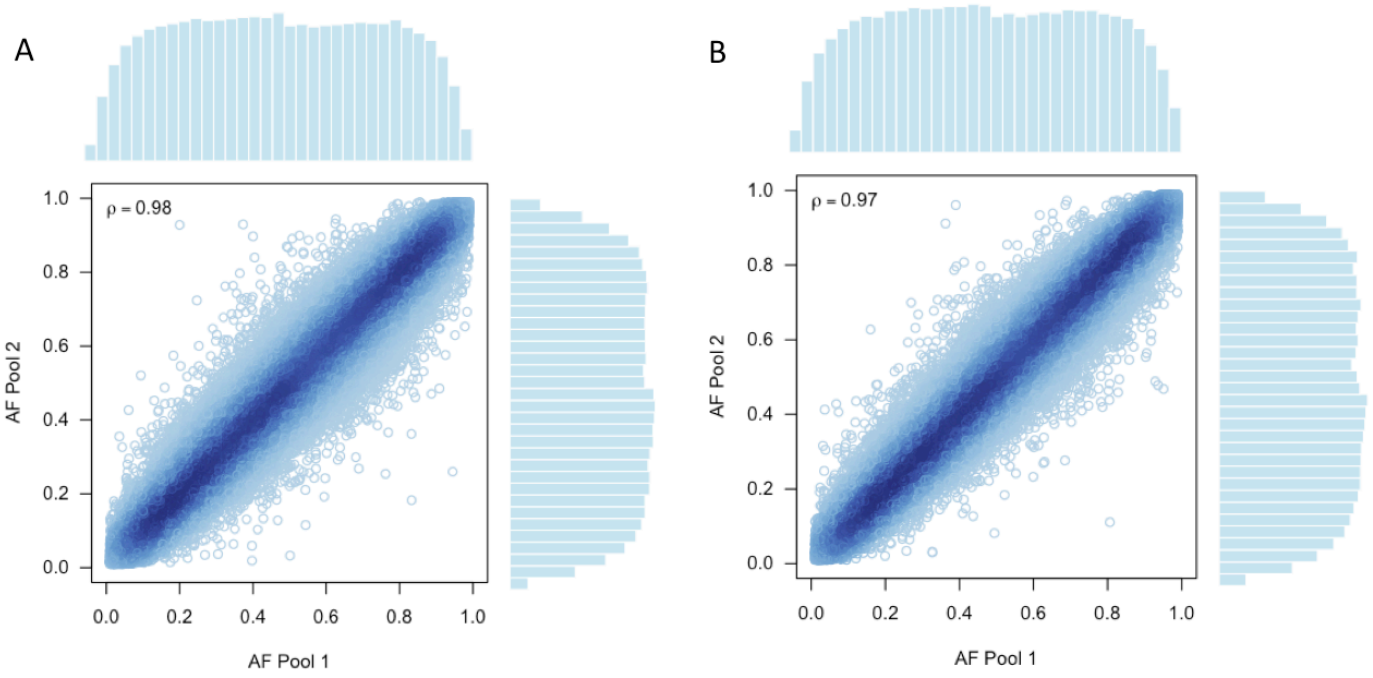
If (3) and (4) are combined we have:

$$t_c = -\log(-\log(1-\alpha)) \qquad (5)$$

Thus far, those *P*-values for which the transformation $h(x)=-\log(-\log(1-x))$ is greater than (4) are said to be significant. We implemented this procedure in R (R Development Core Team, 2011), considering different scenarios, e.g. different n values for cases and controls, markers and number of steps, and it was applied to the *P*-values generated by our *pb*GWAS strategy. Results are presented in Supplementary Figures 4-8, and Supplementary Table 1.

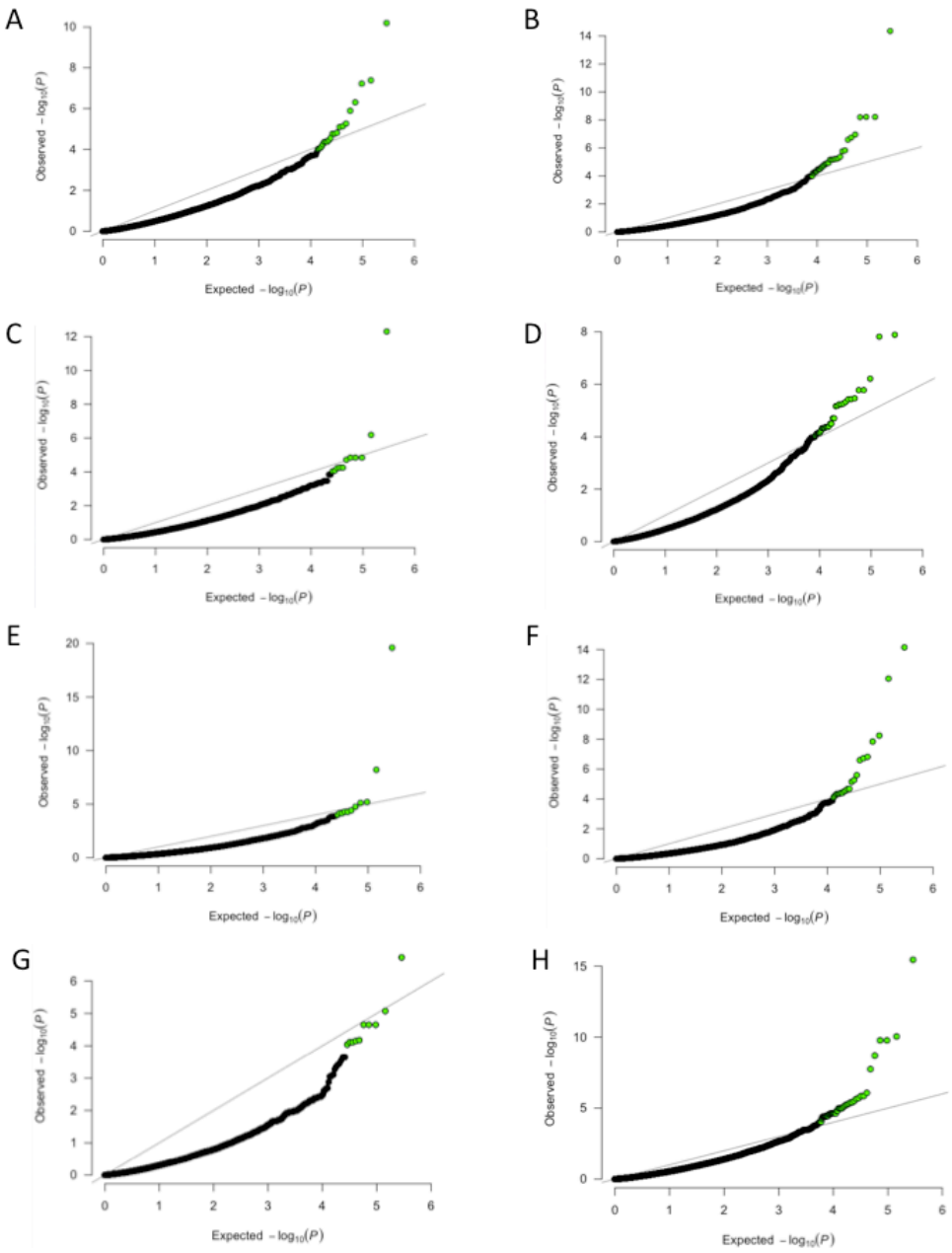
SUPPLEMENTARY FIGURES.

Supplementary Figure 1



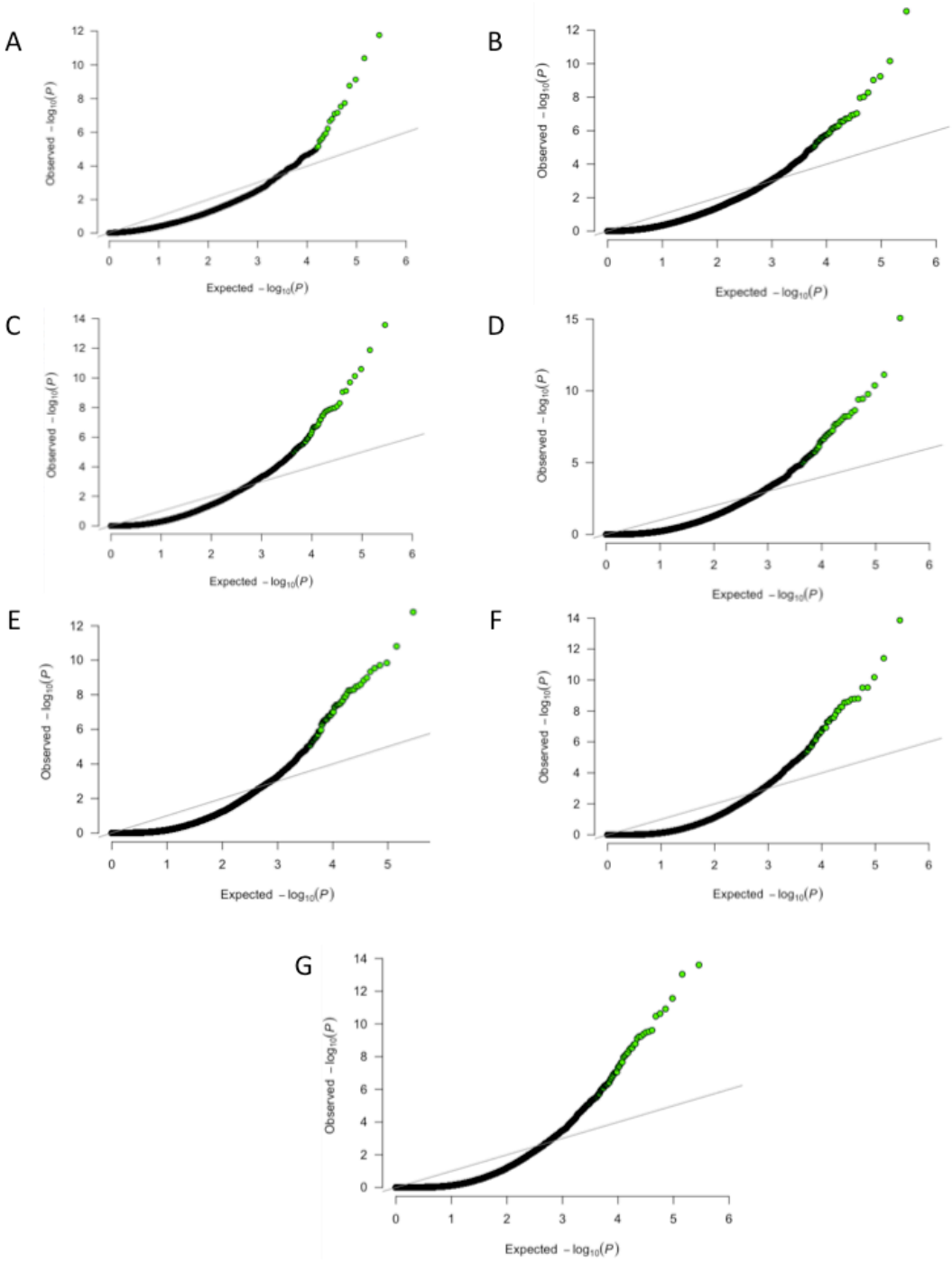
Scatter plots, correlation analyses, and histograms for the allele frequencies (AFs) obtained for two technical replicates using our *pb*GWAS strategy. Panel **A** depicts the results for the group of cases and **B** for the controls. In there, dots represent estimated AFs for each SNP; the *x*-axis corresponds to the AFs for the first replicate and the *y*-axis for the second replicate. Vertical (top) and horizontal (right) bars correspond to the histograms for the AFs in Pools 1 and 2, respectively. Comparison of the AF density distribution functions within each group using R (R Development Core Team, 2011) and the `sm` package (Bowman & Azzalini, 2010) with *B*=100 replicates shows that these are statistically equivalent (cases: *P*=0.36; controls: *P*=0.11).

Supplementary Figure 2



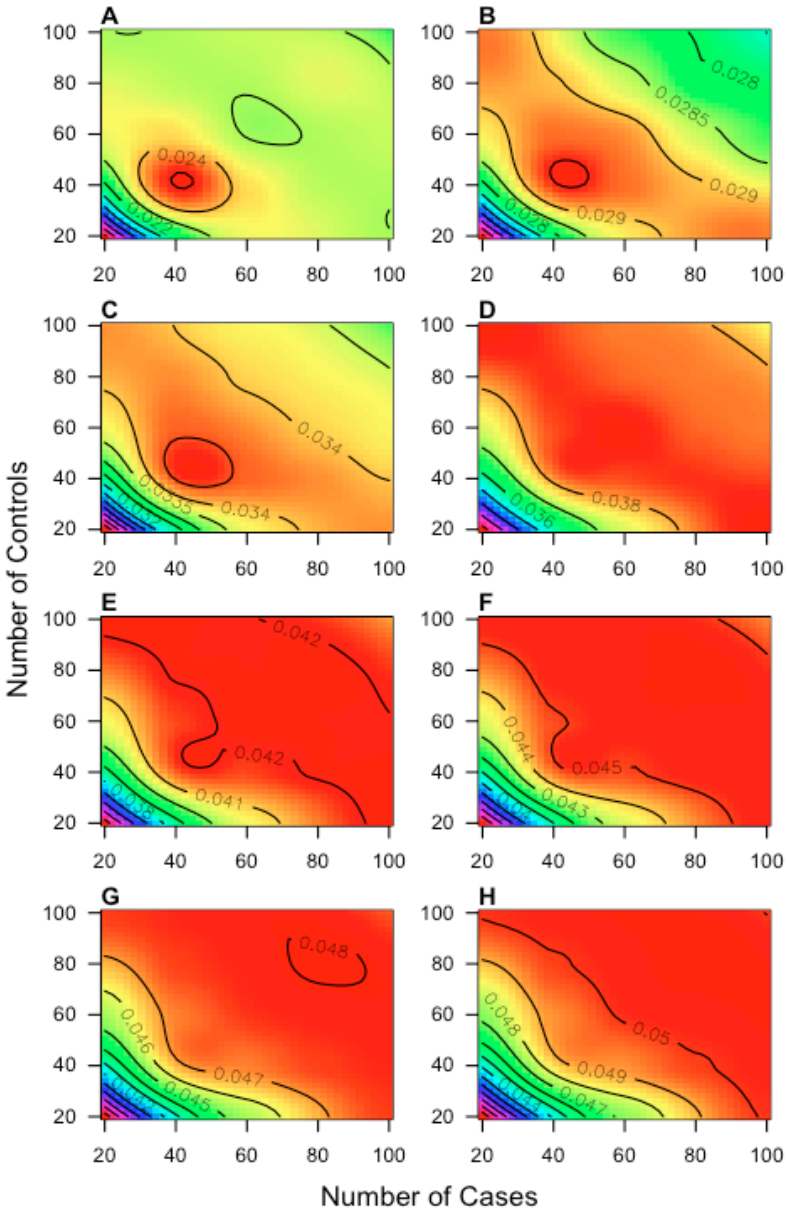
Quantile-quantile plots for observed versus expected FDR-corrected $-\log_{10}(P)$ values for each of eight pairs (from **A** to **H**, respectively) of DNA pools generated via bootstrap as described in our *pb*GWAS strategy. In these plots, dots represent the $-\log_{10}(P)$ values for 287,368 single nucleotide polymorphisms (SNPs); green dots correspond to those SNPs for which the $-\log_{10}(P)$ is greater than four, e.g., $P<10^{-4}$.FDR = False Discovery Rate.

Supplementary Figure 3



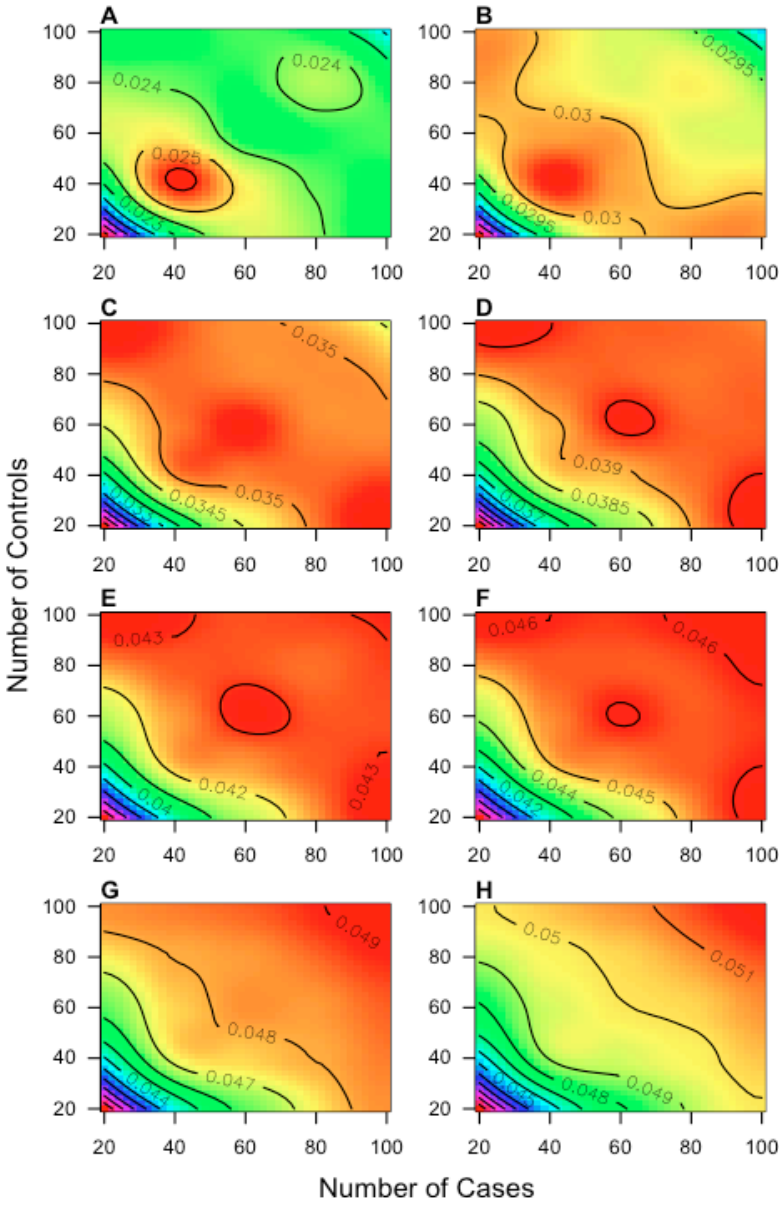
Quantile-quantile plots for observed versus expected FDR-corrected $-\log_{10}(P)$ values after combining the *P*-values from **(A)** steps 1 to 2, **(B)** 1 to 3, **(C)** 1 to 4, **(D)** 1 to 5, **(E)** 1 to 6, **(F)** 1 to 7 and **(G)** 1 to 8 using the Stouffer's method as described in our *pb*GWAS strategy (Figure 1). Abbreviations and conventions as in Supplementary Figure 2.

Supplementary Figure 4



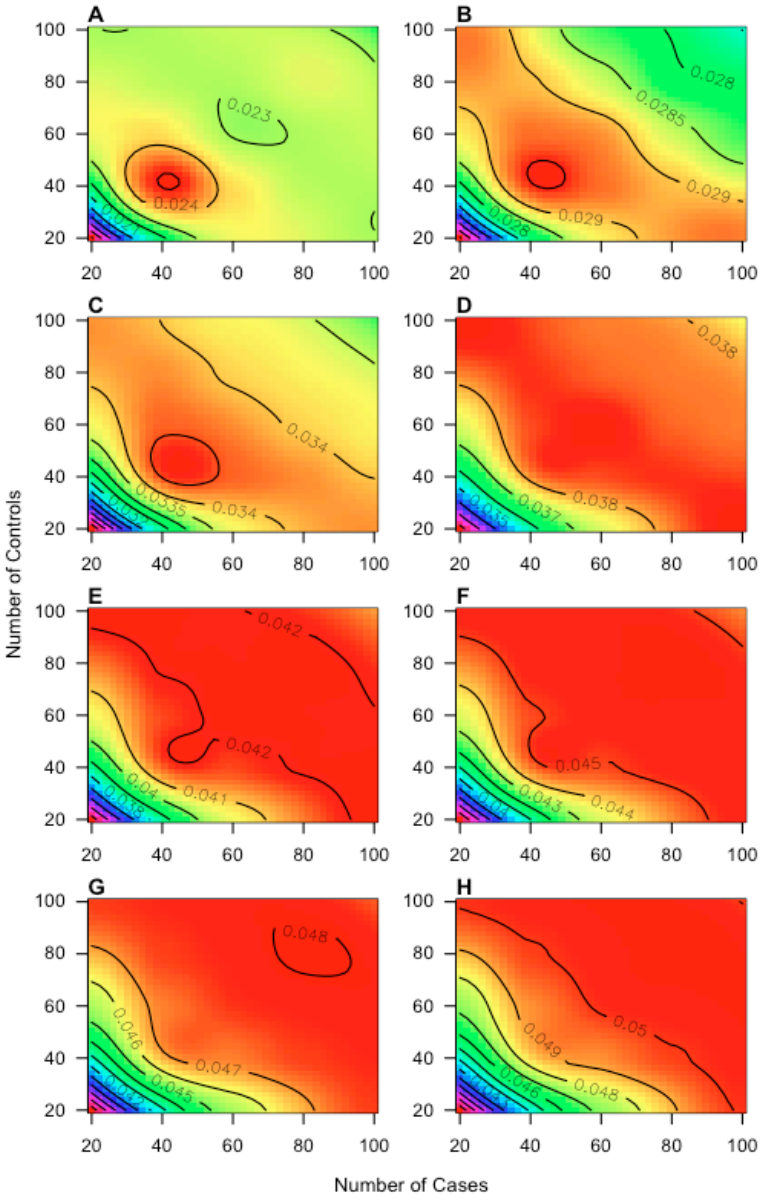
Contour plots for rejection rates of H_0 for allele frequencies that do not differ between cases and controls when a *pb*GWAS strategy is used; *m*=1,000 SNPs. Stouffer's method was used to combine the *P*-values (see Materials and Methods and Figure 1 of the main manuscript) from **(A)** step 1, **(B)** 1 to 2, **(C)** 1 to 3, **(D)** 1 to 4, **(E)** 1 to 5, **(F)** 1 to 6, **(G)** 1 to 7 and **(H)** 1 to 8. The *x* and *y* axes represent the total number of DNA samples available from cases and controls, respectively. The type I error probability was fixed at α=0.05.. High rejection rates are represented in red.

Supplementary Figure 5



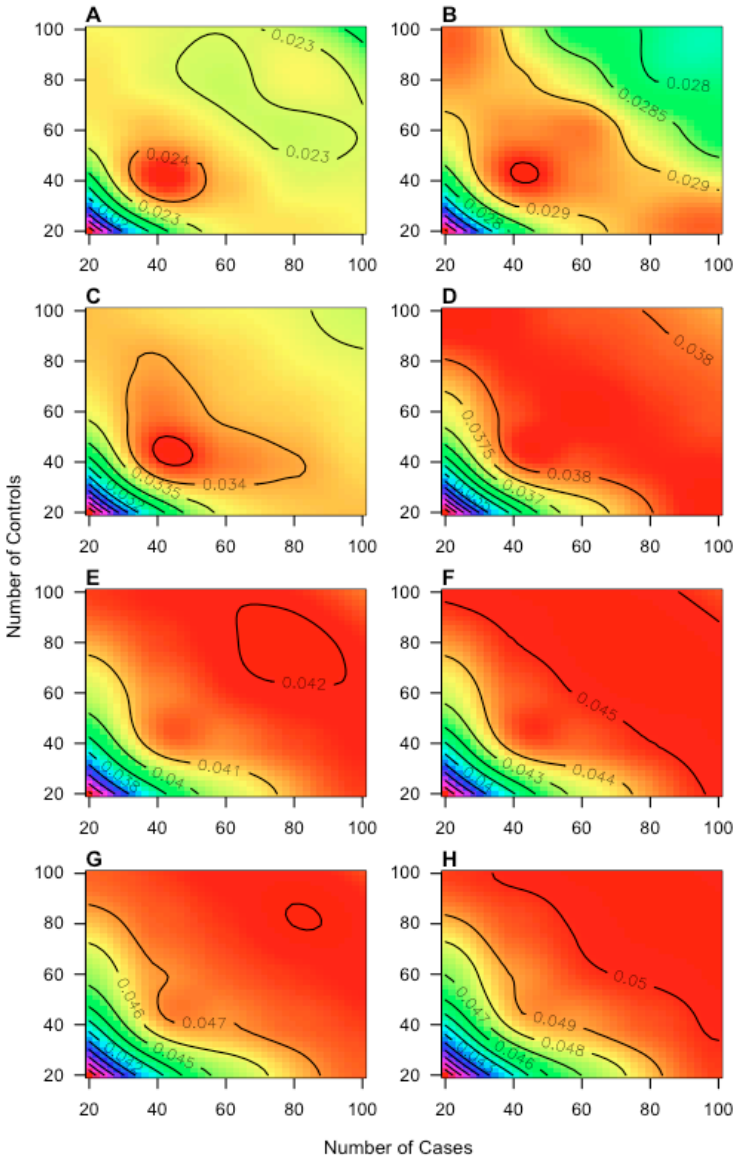
Contour plots for rejection rates of H_0 for allele frequencies that do not differ between cases and controls when a *pb*GWAS strategy is used; *m*=10,000 SNPs. Stouffer's method was used to combine the *P*-values (see Materials and Methods and Figure 1 of the main manuscript) from **(A)** step 1, **(B)** 1 to 2, **(C)** 1 to 3, **(D)** 1 to 4, **(E)** 1 to 5, **(F)** 1 to 6, **(G)** 1 to 7 and **(H)** 1 to 8. The *x* and *y* axes represent the total number of DNA samples available from cases and controls, respectively. The type I error probability was fixed at α=0.05. High rejection rates are represented in red.

Supplementary Figure 6



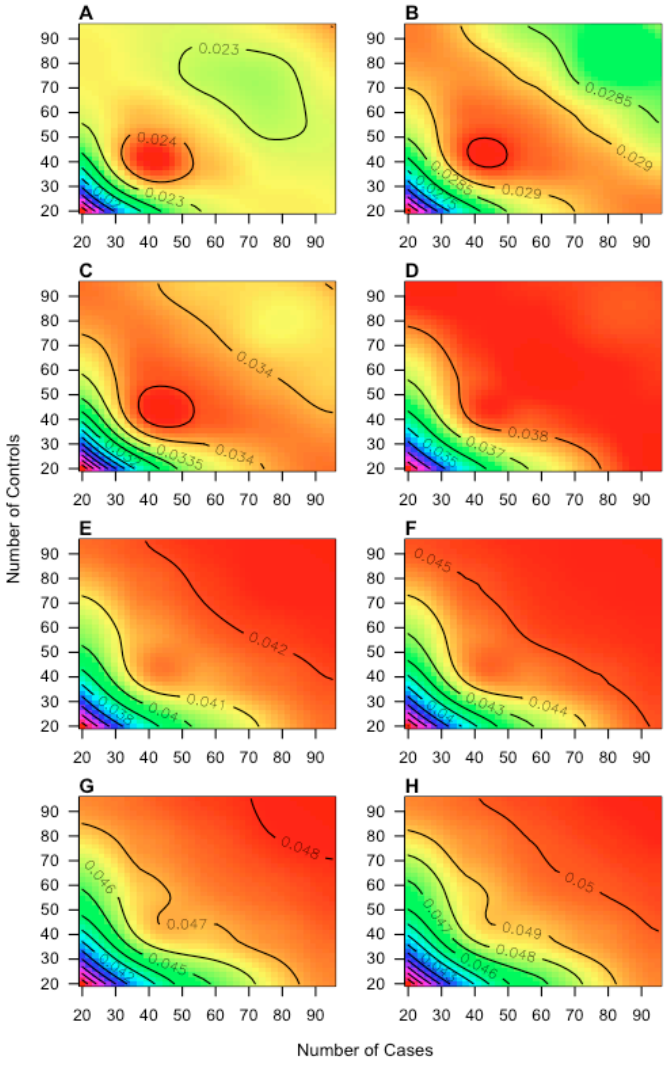
Contour plots for rejection rates of H_0 for allele frequencies that do not differ between cases and controls when a *pb*GWAS strategy is used; *m*=100,000 SNPs. Stouffer's method was used to combine the *P*-values (see Materials and Methods and Figure 1 of the main manuscript) from **(A)** step 1, **(B)** 1 to 2, **(C)** 1 to 3, **(D)** 1 to 4, **(E)** 1 to 5, **(F)** 1 to 6, **(G)** 1 to 7 and **(H)** 1 to 8. The *x* and *y* axes represent the total number of DNA samples available from cases and controls, respectively. The type I error probability was fixed at α=0.05. High rejection rates are represented in red.

Supplementary Figure 7



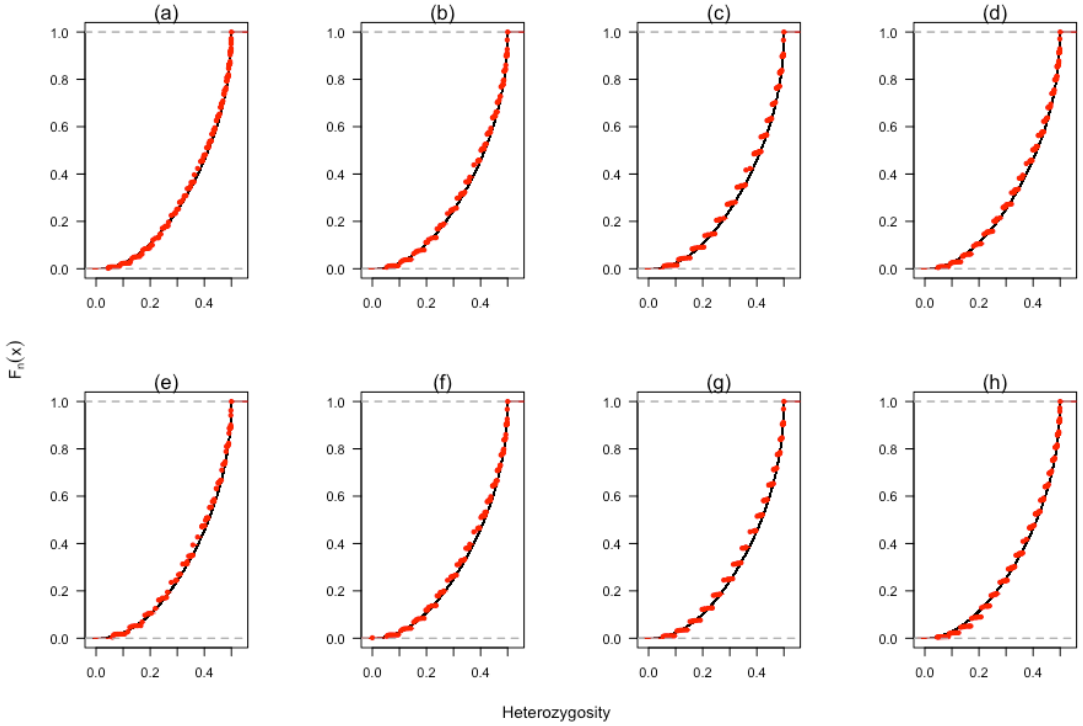
Contour plots for rejection rates of H_0 for allele frequencies that do not differ between cases and controls when a *pb*GWAS strategy is used; *m*=300,000 SNPs. Stouffer's method was used to combine the *P*-values (see Materials and Methods and Figure 1 of the main manuscript) from **(A)** step 1, **(B)** 1 to 2, **(C)** 1 to 3, **(D)** 1 to 4, **(E)** 1 to 5, **(F)** 1 to 6, **(G)** 1 to 7 and **(H)** 1 to 8. The *x* and *y* axes represent the total number of DNA samples available from cases and controls, respectively. The type I error probability was fixed at α=0.05. High rejection rates are represented in red.

Supplementary Figure 8



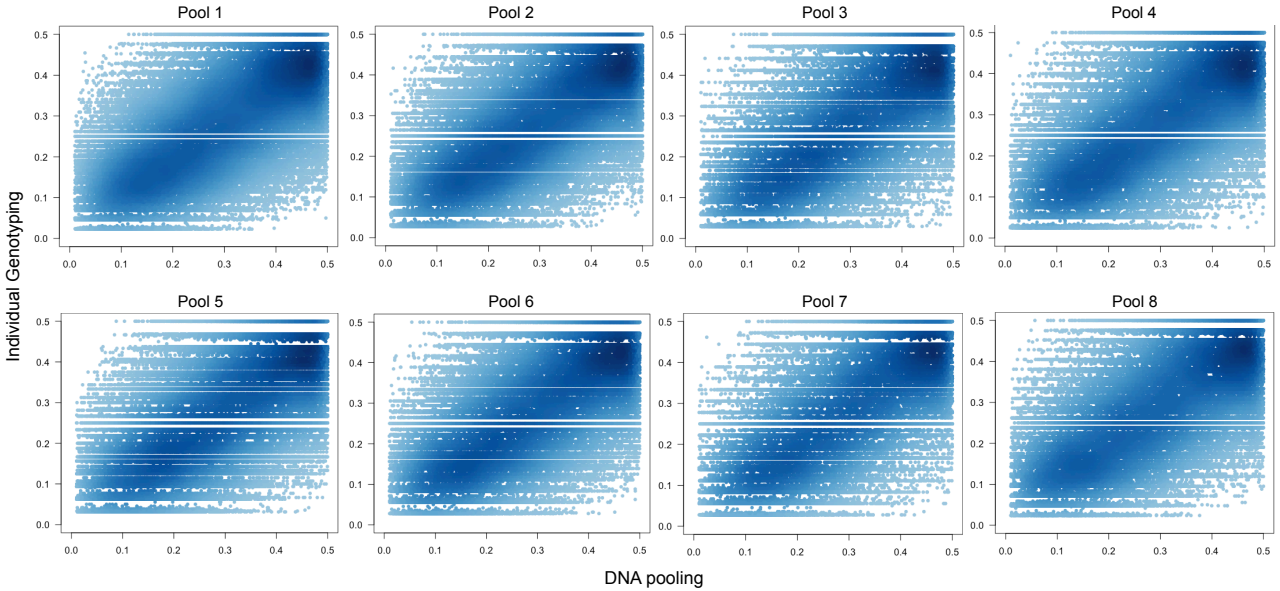
Contour plots for rejection rates of H_0 for allele frequencies that do not differ between cases and controls when a *pb*GWAS strategy is used; *m*=500,000 SNPs. Stouffer's method was used to combine the *P*-values (see Materials and Methods and Figure 1 of the main manuscript) from **(A)** step 1, **(B)** 1 to 2, **(C)** 1 to 3, **(D)** 1 to 4, **(E)** 1 to 5, **(F)** 1 to 6, **(G)** 1 to 7 and **(H)** 1 to 8. The *x* and *y* axes represent the total number of DNA samples available from cases and controls, respectively. The type I error probability was fixed at α=0.05. High rejection rates are represented in red.

Supplementary Figure 9.



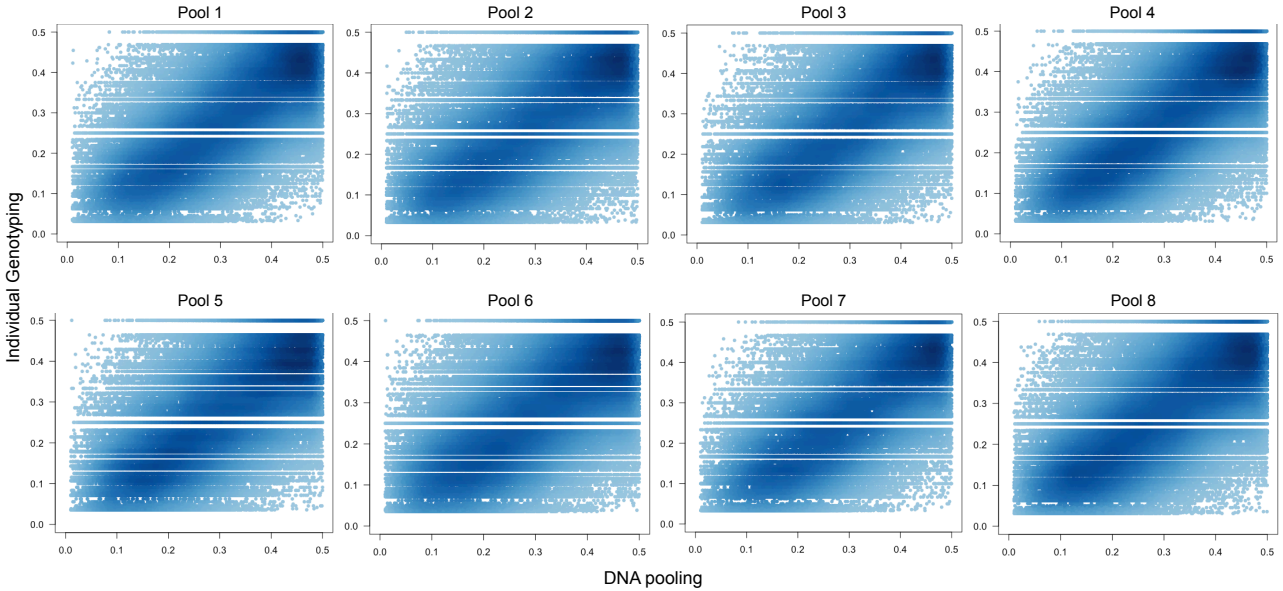
Cumulative distribution function of the heterozygosity values in cases and controls for individual genotyping (red dots) and DNA pooling (black dots).

Supplementary Figure 10.



Pattern of correlation that was found between gene frequencies estimated by DNA pooling and defined by individual genotyping for cases.

Supplementary Figure 11.



Pattern of correlation that was found between gene frequencies estimated by DNA pooling and defined by individual genotyping for controls.

Supplementary Table 1. Simulation-based H_0 rejections rates when no difference in the allele frequencies between cases and controls is present. For calculation purposes,Type I error probability was fixed at α=0.05..

Number of SNPs	DNA samples		Rejection Rate of H_0 in *pb*GWAS							
	Cases	Controls	S1	S1-2	S1-3	S1-4	S1-5	S1-6	S1-7	S1-8
100,000	20	20	0.0123	0.0238	0.0268	0.0300	0.0316	0.0335	0.0346	0.0353
100,000	25	25	0.0241	0.0298	0.0337	0.0365	0.0396	0.0409	0.0426	0.0437
100,000	30	30	0.0208	0.0260	0.0308	0.0355	0.0386	0.0411	0.0431	0.0445
100,000	35	35	0.0285	0.0306	0.0352	0.0388	0.0423	0.0448	0.0468	0.0484
100,000	40	40	0.0209	0.0299	0.0344	0.0373	0.0400	0.0430	0.0450	0.0471
100,000	45	45	0.0266	0.0290	0.0341	0.0384	0.0423	0.0456	0.0476	0.0494
100,000	50	50	0.0208	0.0271	0.0344	0.0373	0.0409	0.0440	0.0465	0.0488
100,000	55	55	0.0251	0.0310	0.0347	0.0399	0.0492	0.0460	0.0485	0.0509
300,000	20	20	0.0123	0.0241	0.0271	0.0301	0.0319	0.0336	0.0346	0.0354
300,000	25	25	0.0241	0.0299	0.0337	0.0363	0.0395	0.0408	0.0425	0.0436
300,000	30	30	0.0208	0.0265	0.0313	0.0357	0.0385	0.0408	0.0426	0.0441
500,000	20	20	0.0123	0.0234	0.0265	0.0296	0.0316	0.0334	0.0348	0.0356
500,000	25	25	0.0241	0.0299	0.0336	0.0363	0.0395	0.0408	0.0425	0.0436

Supplementary Table 2. number of DNA samples used in each step of the *pb*GWAS strategy that also were individually genotyped

Phenotype	Pool							
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**
EOAD	18	18	17	20	16	18	18	21
LOAD	15	15	16	16	14	14	15	16
Total	33	33	33	36	30	32	33	37

Supplementary Table 3. Estimated linear correlation coefficients (ρ) and 95% confidence intervals (CI) when the heterozygosity values obtained with DNA pooling and individual genotyping for EOAD and LOAD patients are plotted against each other.

Pool	Cases		Controls	
	ρ	95% CI	ρ	95% CI
1	0.6434	(0.6412, 0.6456)	0.6412	(0.6390, 0.6433)
2	0.6480	(0.6459, 0.6502)	0.6130	(0.6106, 0.6153)
3	0.6352	(0.6330, 0.6374)	0.6407	(0.6385, 0.6429)
4	0.6458	(0.6436, 0.6480)	0.6520	(0.6498, 0.6541)
5	0.6408	(0.6386, 0.6430)	0.6202	(0.6179, 0.6224)
6	0.6515	(0.6493, 0.6536)	0.5994	(0.5970, 0.6017)
7	0.6558	(0.6537, 0.6579)	0.6448	(0.6426, 0.6469)
8	0.6393	(0.6371, 0.6415)	0.6176	(0.6153, 0.6199)

REFERENCES

1. Bowman, A. W. and Azzalini, A. (2010). R package 'sm': nonparametric smoothing methods (version 2.2-4) URL http://www.stats.gla.ac.uk/~adrian/sm, http://azzalini.stat.unipd.it/Book_sm

2. Casella, G. and Berger, R.L. (2001). *Statistical Inference*. Duxbury Press.

3. Devroye, L. (1986). *Non-Uniform Random Variate Generation*. New York: Springer-Verlag.

4. Murdoch, D.J., Tsai, Y.L., Adcock, J. (2008). *P*-values are random variables. *The American Statistician*, August 2008, Vol. 62, No. 3, pp. 242-245.

5. R Development Core Team (2011). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org/

6. Sackrowitz, H. and Samuel-Cahn, E. (1999). *P*-values as Random Variables-Expected *P*-values. *The American Statistician*, Vol. 53, No. 4, pp. 326-331.