Multiple testing with prior information identifies loci for exceptional longevity

Edgar Dobriban[1] Kristen Fortney[2] Stuart K. Kim[3] Art B. Owen[1]

[1]Stanford Statistics [2]Stanford Dev. Bio. & Genetics [2]BioAge Labs

Background

- Multiple testing: test the effect of a large number of genetic variants (drugs, interventions)
- To avoid false positives, set stringent significance levels - costly, reduces power
- Can benefit from independent prior information about the effects.
- Example: Two-stage testing

Example: Human Longevity

- Human lifespan has a heritable component of 25-30 %.
- Only one DNA polymorphysm (APOE/TOMM40) replicably associated with exceptional longevity.
- Typical study: GWAS. Compare centenarians vs controls. Test 500K SNPs.
- Challenges: Few centenarians. Small effects of many SNPs.
- Large Prior GWAS: coronary artery disease n >20K, Type II diabetes, etc.

P-value Weighting

Weighted Bonferroni Method (rev. in [5])

- reject H_i if $P_i \le qw_i$
- weights $w_i \ge 0$, $\Sigma_{i=1}^J w_i = J$
- controls FWER at $\alpha := Jq$

Spjøtvoll weights optimize expected number of discoveries [6, 1]:

$$\max_{w\in\mathbb{R}^J} \sum_{i=1}^J \mathbb{P}_{H_i=1}(P_i \le qw_i)$$

$$\text{s.t. } w_i \ge 0, \sum_{i=1}^J w_i = J.$$

requires pilot estimates of effects

Acknowledgements

AFAR/EMF, NIH/NIA, NSF, HHMI.

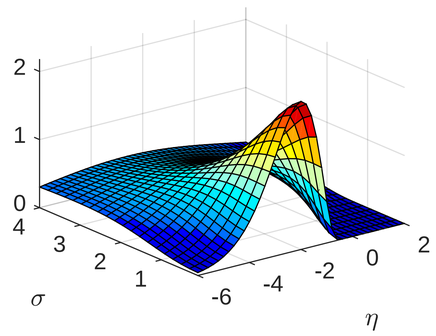

Figure 1: Bayes weights as a function of prior mean and variance.

New Bayes weights

- prior information about effects μ_i: $\mu_i \sim \mathcal{N}(\eta_i, \sigma_i^2)$.
- maximize expected discoveries [3]:

$$\max_{w\in\mathbb{R}^J} \mathbb{E}_\mu \mathbb{E}_T R(w)$$

$$\text{s.t. } w_i \ge 0, \sum_{i=1}^J w_i = J.$$

- Fig. 1: Weights as function of (η, σ^2).
- scales to large problems (e.g., $J = 2$ million)
- only requires summary statistics
- Simulation (Fig. 2): sparse means $\eta_i \sim \pi_0\delta_m + \pi_1\delta_M$, $m = -10^{-3}$, $M = -2$, $q = 10^{-2}$, $\pi_1 \in [0, 0.1]$. $\sigma_i = \sigma \in \{0, 1\}$.
- Bayes weights improve power wrt unweighted, have a worst-case advantage wrt Spjøtvoll weights.

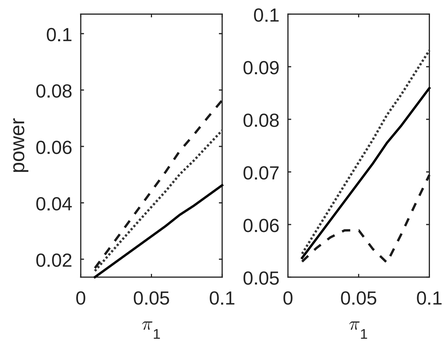

Figure 2: Deterministic (left) and average (right) power as a function of the proportion of large means π_1: Unweighted (solid); Spjøtvoll (dashed); Bayes (dotted).

Discoveries

In [4], discovered four SNPs associated with exceptional longevity

- prior information: GWAS on 8 diseases and 13 traits (Fig. 3)
- studies: New England Centenarian Study (NECS), largest centenarian GWAS (n= 801)
- 90Plus: n=5406 over age 90
- pipeline (Fig. 4)

Software

R package pweight on CRAN
MATLAB at `github.com/dobriban`

Longevity Data Analysis

Disease	Acronym	# Cases	Reference
Age-related macular degeneration	AMD	17,000	[17]
Type 2 diabetes	DIA	34,000	[68]
Rheumatoid arthritis	ART	5,500	[69]
Chronic kidney disease	CKD	5,807	[70]
Late-onset Alzheimer's disease[a]	LOAD	8,000	[29]
Coronary artery disease	CVD	22,233	[71]
Pancreatic cancer	PANC	3,800	[72]
Lung cancer	LUNG	14,900	[73]
Trait	**Acronym**	**# Samples**	**Reference**
Bone density	BMD	32,961	[74]
High-density lipoprotein[b]	HDL	100,184	[19]
Low-density lipoprotein[b]	LDL	100,184	[19]
Triglycerides[b]	TRIG	100,184	[19]
Total cholesterol	TC	100,184	[19]
Body mass index	BMI	123,865	[75]
Waist-hip ratio[b]	WHR	123,865	[75]
Systolic blood pressure[a,b]	SBP	69,395	[76]
Diastolic blood pressure[a]	DBP	69,395	[76]
Beta-cell function[b]	HOMAB	46,186	[77]
Insulin resistance[b]	HOMAIR	46,186	[77]
Fasting insulin	INS	46,186	[77]
Adiponectin	ADIP	49,981	[78]
Control	**Acronym**	**# Samples**	**Reference**
Schizophrenia	SCZ	9394 cases	[79]
Major Depressive Disorder	MDD	9240 cases	[80]
Attention Deficit Hyperactivity Disorder	ADHD	2,064 trios	[81]

[a]Direction of effect was unavailable for these diseases and traits
[b]These traits were excluded from the cross-disease analysis

doi:10.1371/journal.pgen.1005728.t001

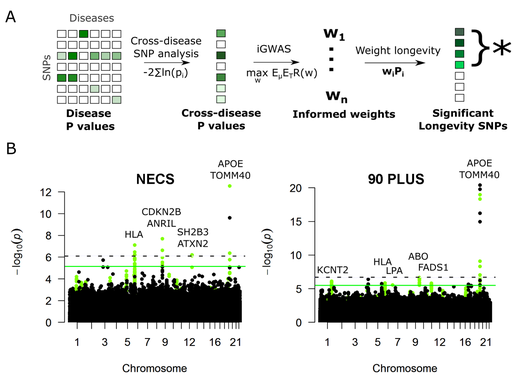

Figure 4: Pipeline.

Replicated discoveries

Source	SNP	Candidate Gene(s)	Combined P[a]
NECS	rs2075650	*TOMM40/APOE*	**2.40E-13**
NECS	rs4977756	*CDKN2B/ANRIL*	**2.82E-03**
NECS	rs3763305	*HLA*	**ns**
NECS	rs3184504	*SH2B3/ATXN2*	**9.41E-03**
Source	**SNP**	**Candidate Gene(s)**	**Combined P[a]**
90PLUS	rs4420638	*TOMM40/APOE*	
90PLUS	rs514659	*ABO*	**6.55E-03**
90PLUS	rs10737670	*KCNT2*	**ns**
90PLUS	rs12194148	*HLA*	**ns**
90PLUS	rs174555	*FADS1*	**ns**
90PLUS	rs10455872	*LPA*	

A general framework

In [2], proposed a more general framework for multiple testing with prior information

- allows any convex constraints
- monotone+bounded, stratified, smooth weights,
- performs well

The future

- new applications
- new optimality criteria

References

[1] Y. Benjamini and Y. Hochberg. Multiple hypotheses testing with weights. *Scandinavian Journal of Statistics*, 24(3):407–418, 1997.

[2] E. Dobriban. A general convex framework for multiple testing with prior information. *arXiv preprint arXiv:1603.05334*, 2016.

[3] E. Dobriban, K. Fortney, S. K. Kim, and A. B. Owen. Optimal multiple testing under a Gaussian prior on the effect sizes. *Biometrika*, 102(4):753–766, 2015.

[4] K. Fortney, E. Dobriban, P. Garagnani, C. Pirazzini, D. Monti, D. Mari, G. Atzmon, N. Barzilai, C. Franceschi, A. B. Owen, and S. K. Kim. Genome-wide scan informed by age-related disease identifies loci for exceptional human longevity. *PLoS Genet*, 11(12): e1005728, 2015.

[5] K. Roeder and L. Wasserman. Genome-wide significance levels and weighted hypothesis testing. *Statistical Science*, 24(4):398–413, 2009.

[6] E. Spjøtvoll. On the optimality of some multiple comparison procedures. *The Annals of Mathematical Statistics*, 43(2):398–411, 1972.