Divergence in polygenic score predictions among populations

Graham Coop

Evolution and Ecology

UC Davis

Population differentiation of polygenic score predictions under stabilizing selection

Sivan Yair[1] and Graham Coop[1]

[1] Center for Population Biology and Department of Evolution and Ecology, University of California, Davis

To whom correspondence should be addressed: ssyair@ucdavis.edu

GWAS have identified many loci of very small effect
That act in a a mainly additive manner

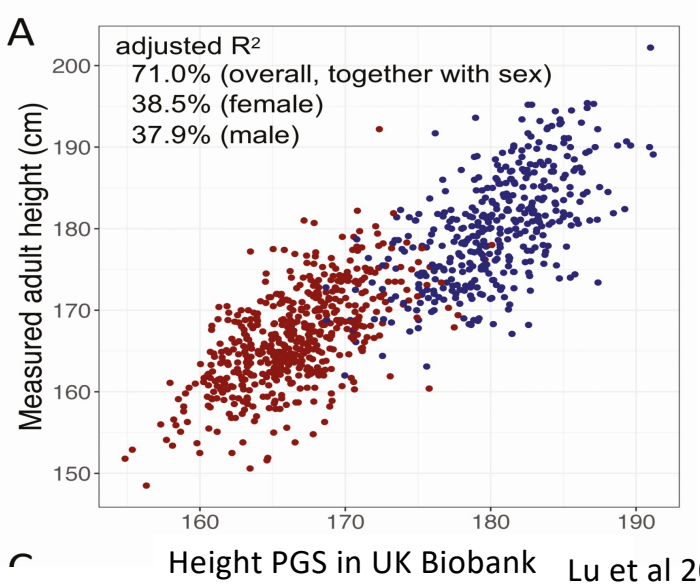
Estimated effect sizes: $\alpha_1, \alpha_2, \alpha_3 \ldots$
Genotypes $g_1, g_2, g_3 \ldots$

Prediction of Total Genetic Value Using Genome-Wide Dense Marker Maps

T. H. E. Meuwissen, B. J. Hayes and M. E. Goddard
GENETICS *April 1, 2001 vol. 157 no. 4 1819-1829*

Polygenic score (PGS) for an individual $Z = \alpha_1 g_1 + \alpha_2 g_2 + \cdots + \alpha_L g_L$

Use in phenotypic predictions
Prediction accuracy of PGS: R^2 of prediction of phenotype by PGS



Lu et al 2021

Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations

Amit V. Khera, Mark Chaffin, Krishna G. Aragam, Mary E. Haas, Carolina Roselli, Seung Hoan Choi, Pradeep Natarajan, Eric S. Lander, Steven A. Lubitz, Patrick T. Ellinor & Sekar Kathiresan

Nature Genetics **50**, 1219–1224 (2018) | Cite this article

Gene discovery and polygenic prediction from a genome-wide association study of educational attainment in 1.1 million individuals

James J. Lee, Robbee Wedow, [...] David Cesarini

Nature Genetics **50**, 1112–1121 (2018) | Cite this article

Utility of polygenic embryo screening for disease depends on the selection strategy

Todd Lencz, Daniel Backenroth, Einat Granot-Hershkovitz, Adam Green, Kyle Gettler, Judy H Cho, Omer Weissbrod, Or Zuk, Shai Carmi

Evidence of widespread selection on standing variation in Europe at height-associated SNPs

Michael C Turchin, Charleston WK Chiang, Cameron D Palmer, Sriram Sankararaman, David Reich, Genetic Investigation of ANthropometric Traits (GIANT) Consortium & Joel N Hirschhorn

Nature Genetics **44**, 1015–1019 (2012) | Cite this article

Are changes in ancient human pops. Genetic or environmental

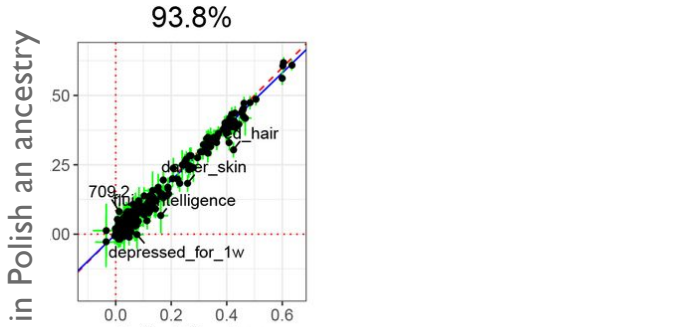
Marciniak *et al*. 2021

See also Cox *et al*. 2021

Polygenic score (PGS) for an individual $Z = \alpha_1 g_1 + \alpha_2 g_2 + \cdots + \alpha_L g_L$

Use in phenotypic predictions

Prediction accuracy of PGS r^2 of prediction of phenotype by PGS



Portability issue

Prediction accuracy

Drops across different groups

Comments

High-resolution portability of 245 polygenic scores when derived and applied in the same cohort

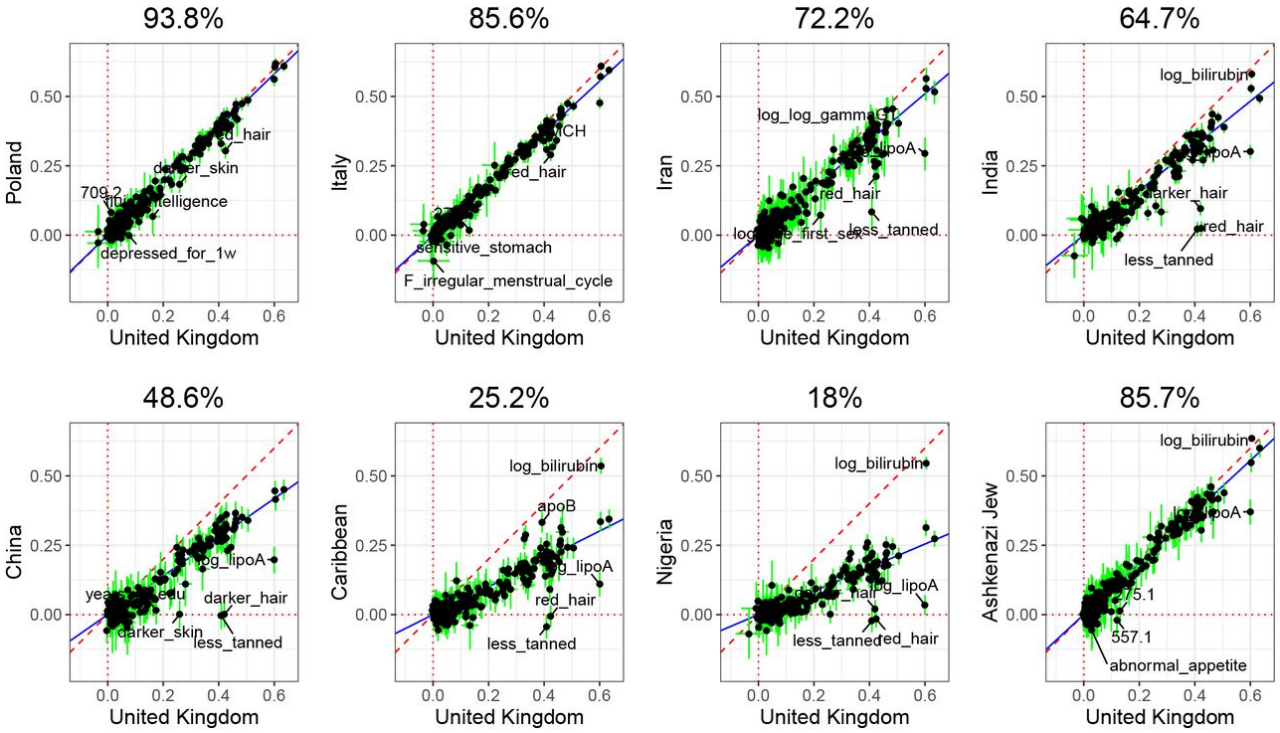
Florian Privé, Hugues Aschard, Shai Carmi, Lasse Folkersen, Clive Hoggart, Paul F. O'Reilly, Bjarni J. Vilhjálmsson

doi: https://doi.org/10.1101/2021.02.05.21251061

Polygenic score (PGS) for an individual $Z = \alpha_1 g_1 + \alpha_2 g_2 + \cdots + \alpha_L g_L$

Use in phenotypic predictions

Prediction accuracy of PGS r^2 of prediction of phenotype by PGS



Portability issue

Prediction accuracy

Drops across different groups

Comments

High-resolution portability of 245 polygenic scores when derived and applied in the same cohort

Florian Privé, Hugues Aschard, Shai Carmi, Lasse Folkersen, Clive Hoggart, Paul F O'Reilly, Bjarni J.Vilhjálmsson

doi: https://doi.org/10.1101/2021.02.05.21251061

Big list of issues in the portability of polygenic scores across groups

1) Population stratification: GWAS effect sizes can be biased absorb the effects of covarying environmental and genetic gradients.

Reduced signal for polygenic adaptation of height in UK Biobank

Jeremy J. Berg[*,†,1], Arbel Harpak[*,1,2], Nasa Sinnott-Armstrong[*,3], Anja Moltke Jørgensen[4], Hakhamanesh Mostafavi[1], Yair Field[3], Evan A Boyle[3], Xinjun Zhang[5], Fernando Racimo[4], Jonathan K Pritchard[†,2,3,6], Graham Coop[†,7]

Signals of polygenic adaptation on height have been overestimated due to uncorrected population structure in genome-wide association studies

Mashaal Sohail[1,2,3+], Robert M. Maier[3,4,5+], Andrea Ganna[3,4,5,6,7], Alexander Bloemendal[3,4,5], Alicia R. Martin[3,4,5], Michael C. Turchin[8,9], Charleston W. K. Chiang[10], Joel N. Hirschhorn[3,11,12], Mark Daly[3,4,5,7], Nick Patterson[3,13], Benjamin Neale[3,4,5*], Iain Mathieson[14*], David Reich[3,13,15*], Shamil Sunyaev[1,2,3*]

Big list of issues in the portability of polygenic scores across groups

1) Population stratification: GWAS effect sizes can be biased absorb the effects of covarying environmental and genetic gradients.

2) We are missing a large number of variants and the genetic architecture will differ somewhat among populations, e.g. due to drift.

3) Effect sizes of the loci we have may not transfer over groups

- We assume no dominance or epistasis.
- Not causal alleles, but rather alleles that are in linkage disequilibrium (LD) with them
- Gene by environment interactions
- Differences in assortative mating and indirect effects

4) Groups of people experience different distributions of environments so can differ in their phenotypic mean

Polygenic score (PGS) for an individual $Z = \alpha_1 g_1 + \alpha_2 g_2 + \cdots + \alpha_L g_L$

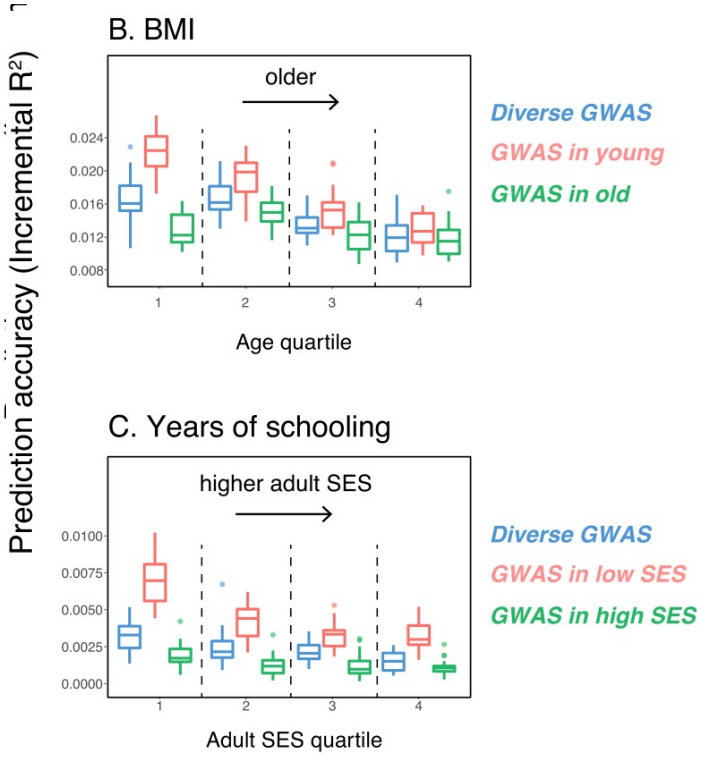
Use in phenotypic predictions

Prediction accuracy of PGS r^2 of prediction of phenotype by PGS

Portability issue

Prediction accuracy

Drops across different groups



› Elife. 2020 Jan 30;9:e48376. doi: 10.7554/eLife.48376.

Variable prediction accuracy of polygenic scores within an ancestry group

Hakhamanesh Mostafavi [# 1], Arbel Harpak [# 1], Ipsita Agarwal [1], Dalton Conley [2 3], Jonathan K Pritchard [4 5 6], Molly Przeworski [1 7]

Big list of issues in the portability of polygenic scores across groups

1) Population stratification: GWAS effect sizes can be biased absorb the effects of covarying environmental and genetic gradients.

2) We are missing a large number of variants and the genetic architecture will differ somewhat among populations, e.g. due to drift.

3) Effect sizes of the loci we have may not transfer over groups

- We assume no dominance or epistasis.
- Not causal alleles, but rather alleles that are in linkage disequilibrium (LD) with them
- Gene by environment interactions
- Differences in assortative mating and indirect effects

4) Groups of people experience different distributions of environments so can differ in their phenotypic mean

The Apportionment of Human Diversity

R. C. LEWONTIN

Committee on Evolutionary Biology, University of Chicago, Chicago, Illinois

1972

Modern version (expressed in terms of F_{ST})

Little variance in genotypes is attributable to differences among human “ancestries” (e.g. continent-level populations)

Most common genetic diversity is present at the level of individual human ‘populations’

Little loss in heterozygosity in sub-populations compared to world wide populations

So can allele frequency changes contribute much to lack of portability?

Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations

Amit V. Khera, Mark Chaffin, Krishna G. Aragam, Mary E. Haas, Carolina Roselli, Seung Hoan Choi, Pradeep Natarajan, Eric S. Lander, Steven A. Lubitz, Patrick T. Ellinor & Sekar Kathiresan

Nature Genetics **50**, 1219–1224 (2018) | Cite this article

Utility of polygenic embryo screening for disease depends on the selection strategy

Gene discovery and polygenic prediction from a genome-wide association study of educational attainment in 1.1 million individuals

To Shai Carmi

James J. Lee, Robbee Wedow, [...] David Cesarini

Nature Genetics **50**, 1112–1121 (2018) | Cite this article

Evidence of widespread selection on standing variation in Europe at height-associated SNPs

Michael C Turchin, Charleston WK Chiang, Cameron D Palmer, Sriram Sankararaman, David Reich, Genetic Investigation of ANthropometric Traits (GIANT) Consortium & Joel N Hirschhorn

Nature Genetics **44**, 1015–1019 (2012) | Cite this article

Prediction
vs
Understanding genetic causes of variation
& Societal implications

Are changes in ancient human pops. Genetic or environmental

Marciniak *et al*. 2021

See also Cox *et al*. 2021

Prediction
vs
Understanding genetic causes of variation
& Societial Implications

The Apportionment of Human Diversity

R. C. LEWONTIN

Committee on Evolutionary Biology, University of Chicago, Chicago, Illinois

1972

19 December 1975, Volume 190, No. 4220 **SCIENCE**

The Heritability Hang-up

The role of variance analysis in human genetics is discussed.

M. W. Feldman and R. C. Lewontin

The nature-nurture issue has provided some of the most keenly contested debates in the fields of biology, psychology, sociology, and physics during the past 5 years. other quantitative traits in humans, are all based on a fundamental methodology that was invented by R. A. Fisher, the analysis of variance. The analysis of variance is

turbation analysis. It is assumed that the actual IQ of an individual is some unknown function of genotype (G) and environment (E)

$$IQ = f(G,E)$$

In any given population, there is some joint distribution of genotypes and environments

$$\phi(G,E)$$

and this joint distribution is mapped onto a distribution of IQ scores $\theta(IQ)$ by a functional equation.

A complete analysis of the causes of variation would involve predicting the changes in the IQ distribution $\theta(IQ)$ from changes in the distribution of genotypes and environments $\phi(G,E)$. However, such an analysis would require that we know the first partial derivatives of the unknown function $f(G,E)$. What we substitute in-

Am J Hum Genet 26:400–411, 1974

Annotation:

The Analysis of Variance and the Analysis of Causes

R. C. LEWONTIN[1]

The evolution of group differences in changing environments

Arbel Harpak, Molly Przeworski

Version 2 Published: January 25, 2021 • https://doi.org/10.1371/journal.pbio.3001072

Reading tea leaves? Polygenic scores and differences in traits among groups

Graham Coop

Tread Lightly Interpreting Polygenic Tests of Selection

John Novembre and Nicholas H. Barton

GENETICS *April 1, 2018 vol. 208 no. 4 1351-1355;*

Interpreting polygenic scores, polygenic adaptation, and human phenotypic differences

Noah A Rosenberg, Michael D Edge, Jonathan K Pritchard, Marcus W Feldman

Evolution, Medicine, and Public Health, Volume 2019, Issue 1, 2019, Pages 26–34,
https://doi.org/10.1093/emph/eoy036
Published: 27 December 2018 **Article history**

Turn over in polymorphism contributes to low prediction accuracy

The Apportionment of Human Diversity

R. C. LEWONTIN 1972

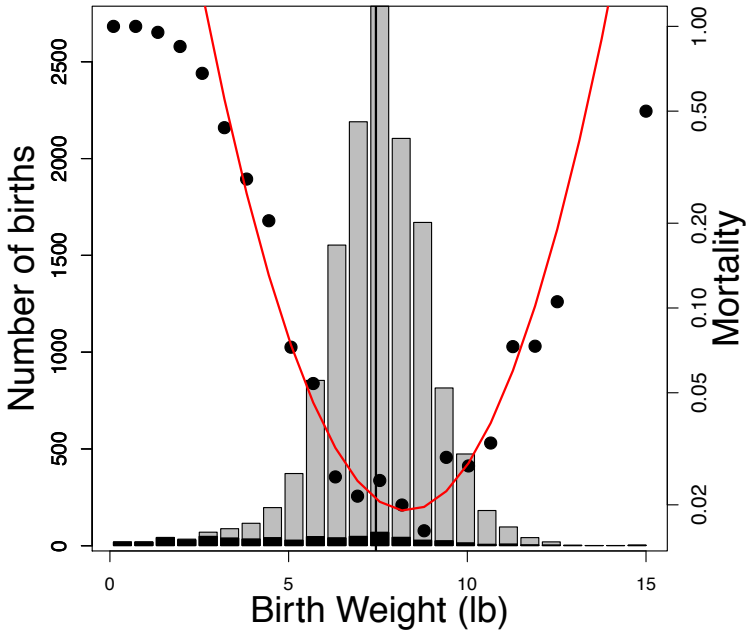
Committee on Evolutionary Biology,
University of Chicago,
Chicago, Illinois

Sharing of alleles among populations

Portability of polygenic scores across populations

Differences in mean polygenic scores

Under models of genetic drift and stabilizing selection



Data from Karn and Penrose (1951)

Turn over in polymorphism contributes to low prediction accuracy

Prediction* accuracy (r^2) correlation between PGS prediction and phenotype
***judged relative to the mean of both in the prediction population**

Reduction in prediction accuracy in moving from genetic values to PGS: $\frac{r^2_{PGS}}{r^2_{GV}}$

Turn over in polymorphism contributes to low prediction accuracy

Prediction* accuracy (r^2) correlation between PGS prediction and phenotype
***judged relative to the mean of both in the prediction population**

Reduction in PGS prediction accuracy : $\frac{r^2_{PGS}}{r^2_{GV}}$

Assume no stratification, no GxE, and that the PGS loci are the causal loci, then reduction is the proportion of genetic variance in a population explained by SNPs in PGS sites

$$\frac{r^2_{PGS}}{r^2_{GV}} = \frac{V_{PGS}}{V_A}$$

Theoretical and empirical quantification of the accuracy of polygenic scores in ancestry divergent populations

Ying Wang, Jing Guo, Guiyan Ni, Jian Yang, Peter M. Visscher & Loic Yengo

Turn over in polymorphism contributes to low prediction accuracy

Prediction* accuracy (r^2) correlation between PGS prediction and phenotype
***judged relative to the mean of both in the prediction population**

Reduction in PGS prediction accuracy : $\frac{r^2_{PGS}}{r^2_{GV}}$

Assume no stratification, no GxE, and that the PGS loci are the causal loci, then reduction is

$$\frac{r^2_{PGS}}{r^2_{GV}} = \frac{V_{PGS}}{V_A}$$

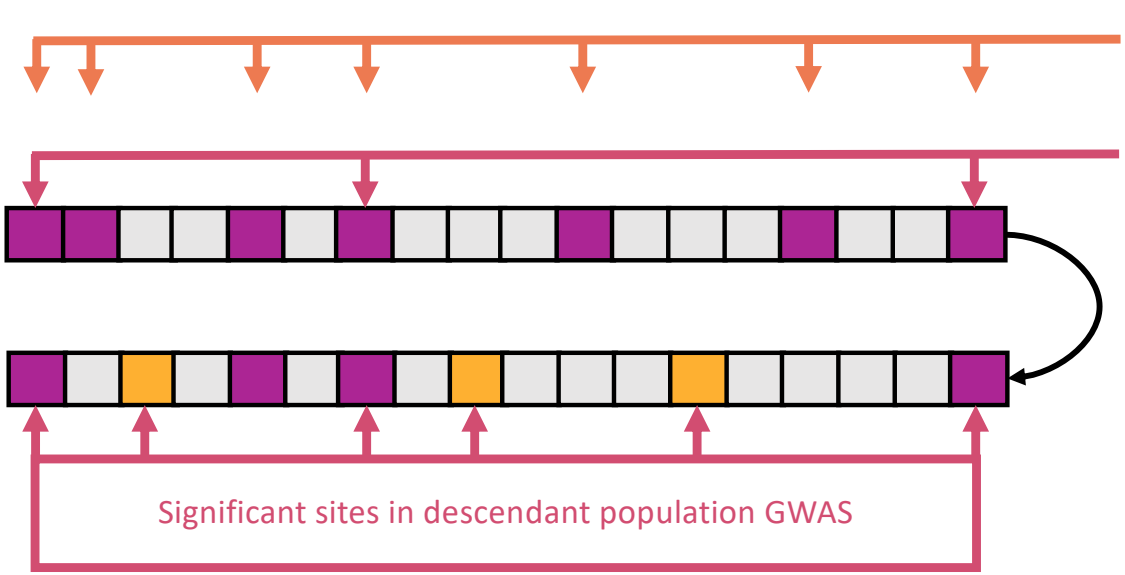
For a set of unlinked SNPs that are uncorrelated (no LD) this is

$$\frac{r^2_{PGS}}{r^2_{GV}} = \frac{V_{PGS}}{V_A} = \frac{\sum_{l \in \mathrm{PGS}} \alpha_l^2 p_l (1 - p_l)}{\sum_{l \in \mathrm{All}} \alpha_l^2 p_l (1 - p_l)}$$

Theoretical and empirical quantification of the accuracy of polygenic scores in ancestry divergent populations

Ying Wang, Jing Guo, Guiyan Ni, Jian Yang, Peter M. Visscher & Loic Yengo ✉

Turn over of polymorphism contribute to low prediction accuracy

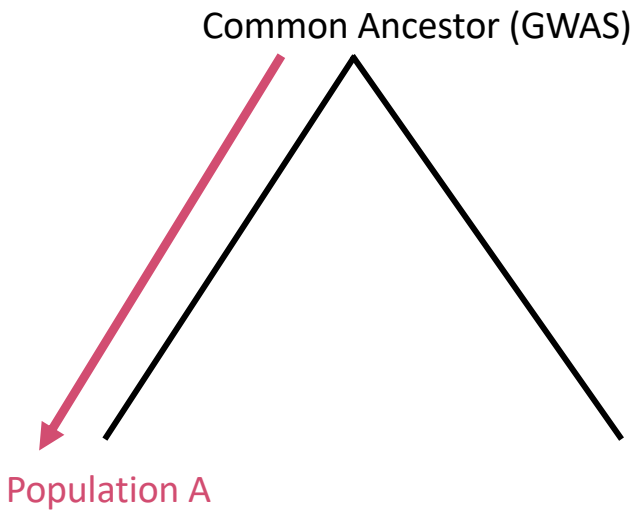


$$\frac{r^2_{PGS}}{r^2_{GV}} = \frac{V_{PGS}}{V_A} = \frac{\sum_{l \in \mathrm{PGS}} \alpha_l^2 p_l (1-p_l)}{\sum_{l \in \mathrm{All}} \alpha_l^2 p_l (1-p_l)}$$

Monomorphic

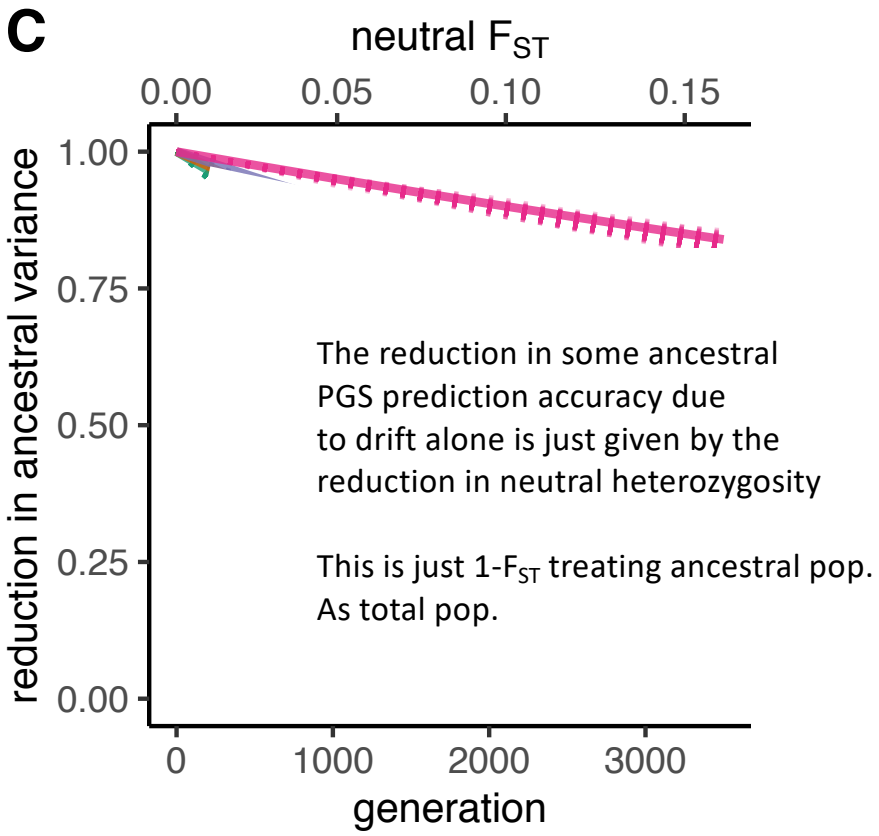
Polymorphism that originated before ancestor

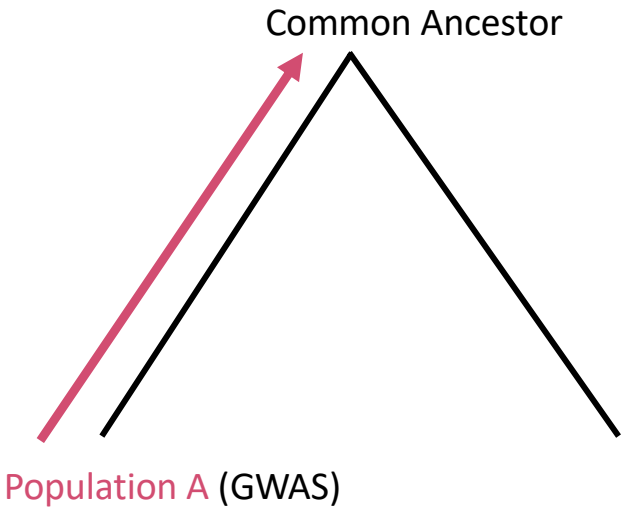
Polymorphism that originated more recently than ancestor



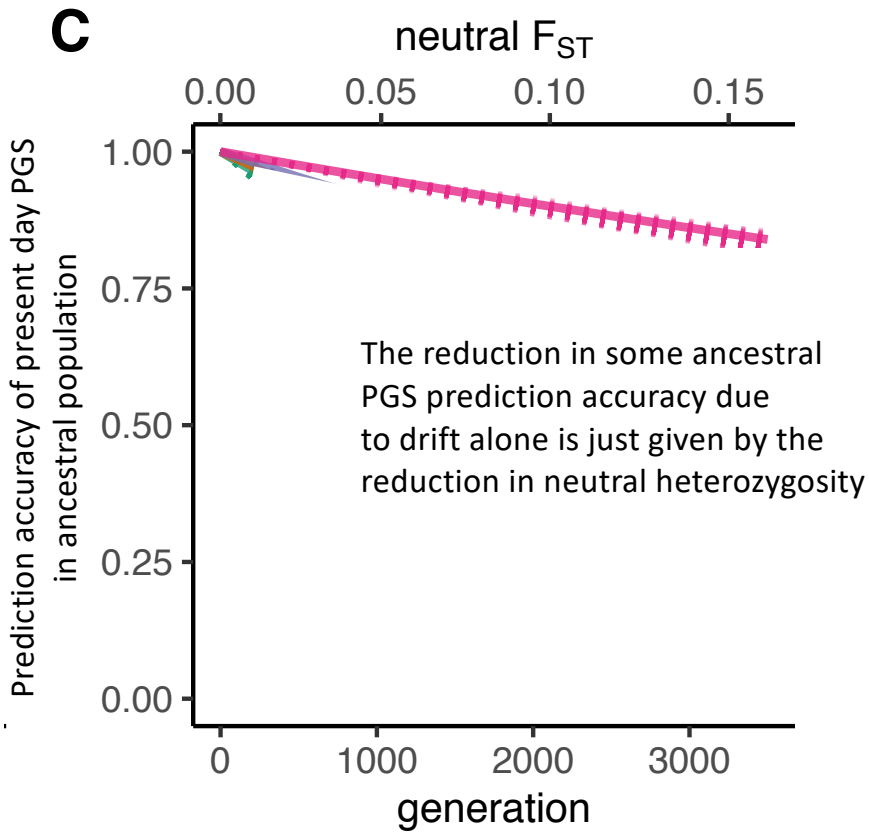
Prediction accuracy (r^2): proportion of variance in a population explained by PGS SNPs

$$\frac{r^2_{PGS}}{r^2_{GV}} = \frac{V_{PGS}}{V_A} = \frac{\sum_{l \in \mathrm{PGS}} \alpha_l^2 p_l (1-p_l)}{\sum_{l \in \mathrm{All}} \alpha_l^2 p_l (1-p_l)}$$

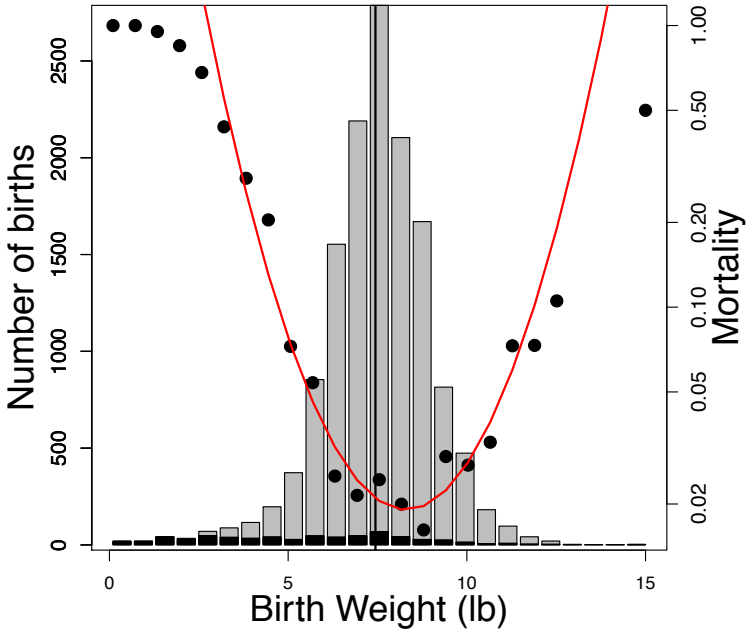




Prediction accuracy (r^2): proportion of variance in a population explained by PGS SNPs



Null model: stabilizing selection on complex traits



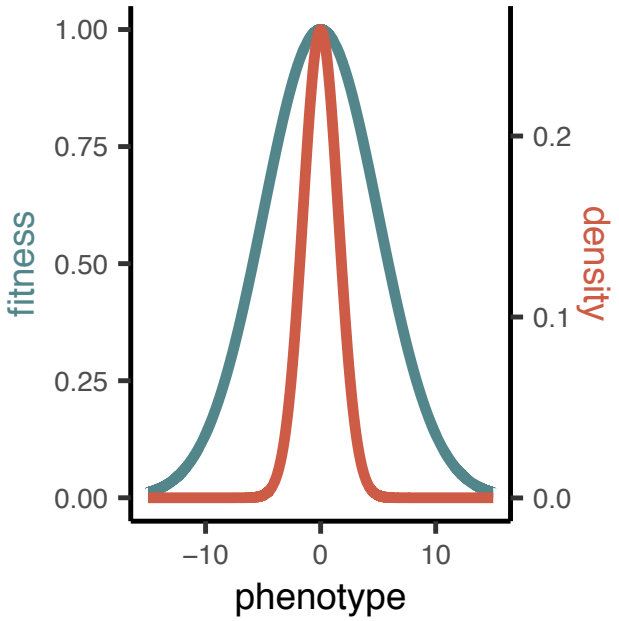
Data from Karn and Penrose (1951)

Evidence of directional and stabilizing selection in contemporary humans

Jaleal S. Sanjak, Julia Sidorenko, Matthew R. Robinson, Kevin R. Thornton, and Peter M. Viss…
+ See all authors and affiliations

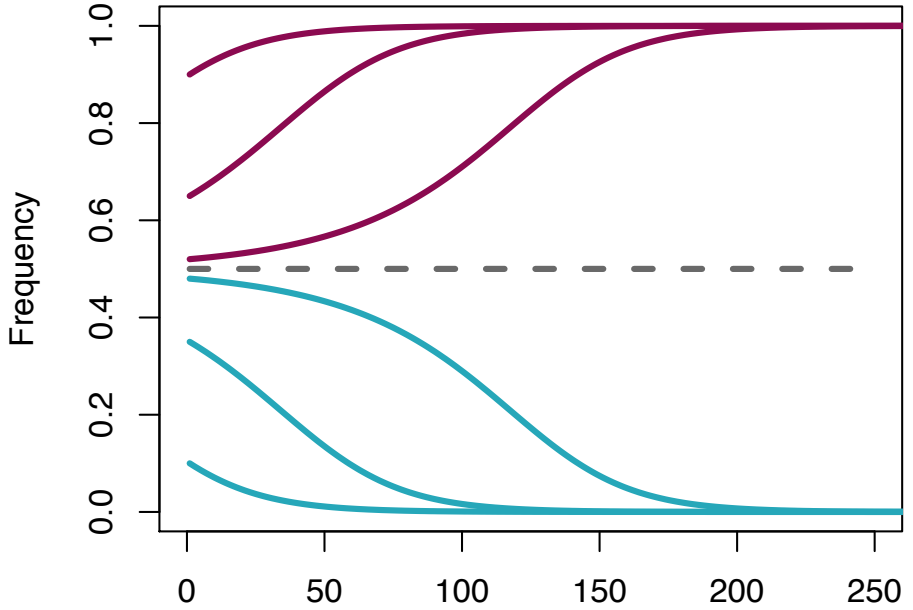
PNAS January 2, 2018 115 (1) 151-156; first published December 18, 2017; https://doi.org/10.1073/pnas.1707227114

Stabilizing selection removes polymorphisms faster than drift

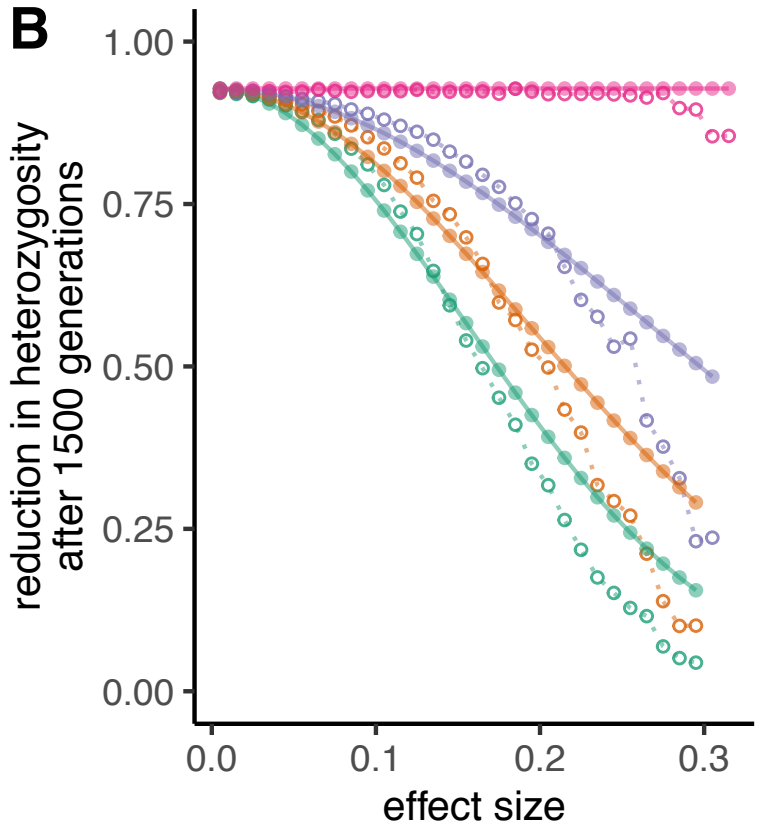


Constant optimum

$$V_A = \sum_{l}^{L} a_l^2 2p_l(1-p_l)$$



Wright 1935



Genet Res. 1988 Aug;52(1):33-43. doi: 10.1017/s0016672300027282.

Quantitative genetic variability maintained by mutation-stabilizing selection balance in finite populations

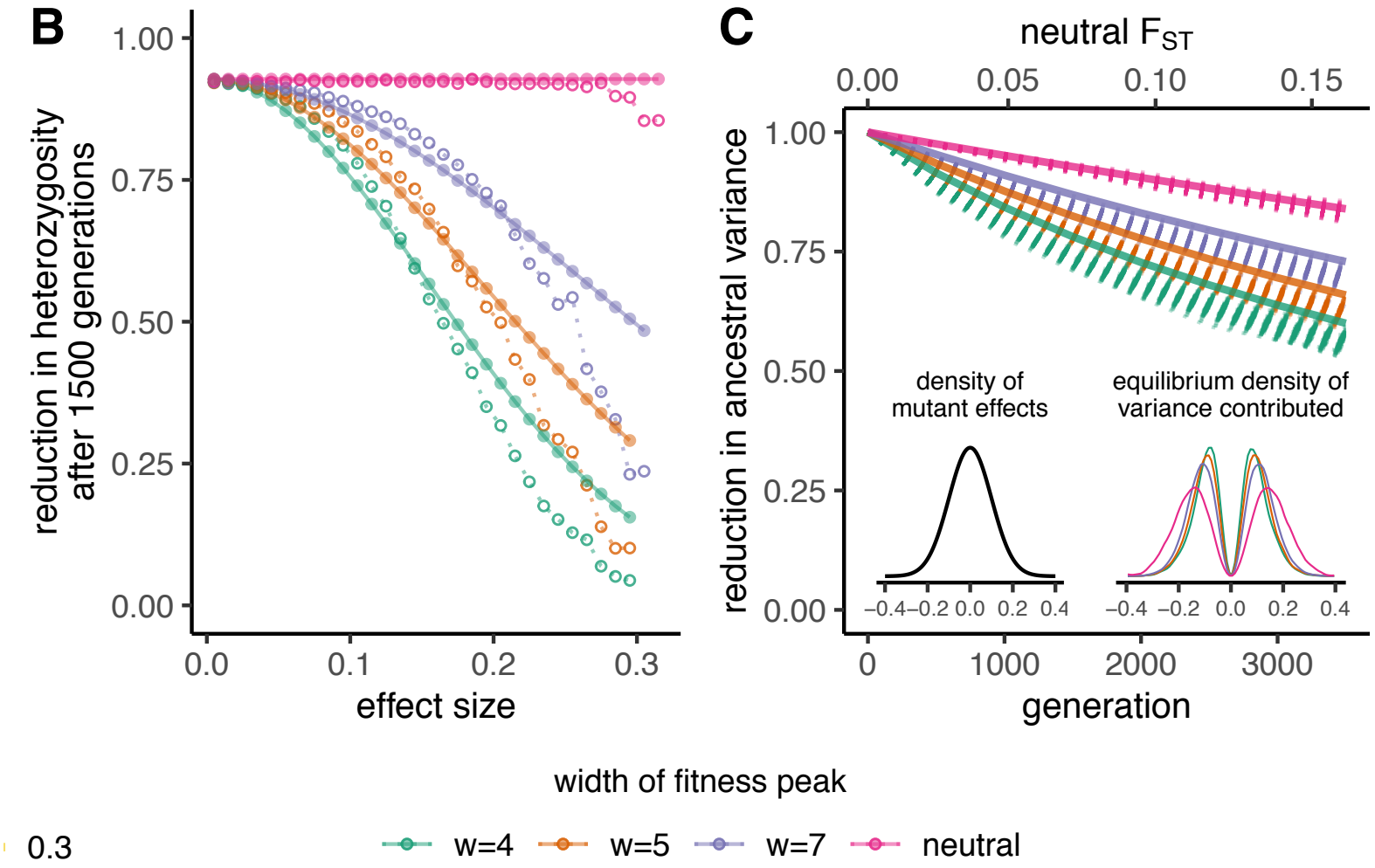
P D Keightley, W G Hill

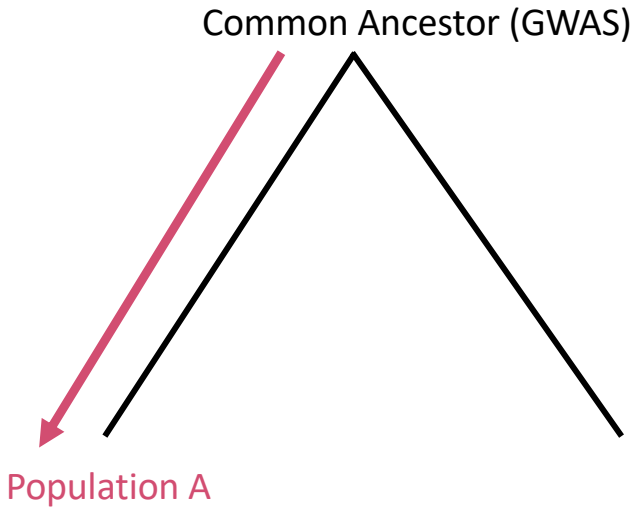
$$\frac{V_{a[\text{anc, desc}]}(\alpha, t)}{V_a(\alpha, 0)} = \left(1 - \frac{1}{2N}\right)^t \times \left(1 - \frac{\alpha^2}{4(w^2 + V_P)}\right)^t$$

width of fitness peak

0.3

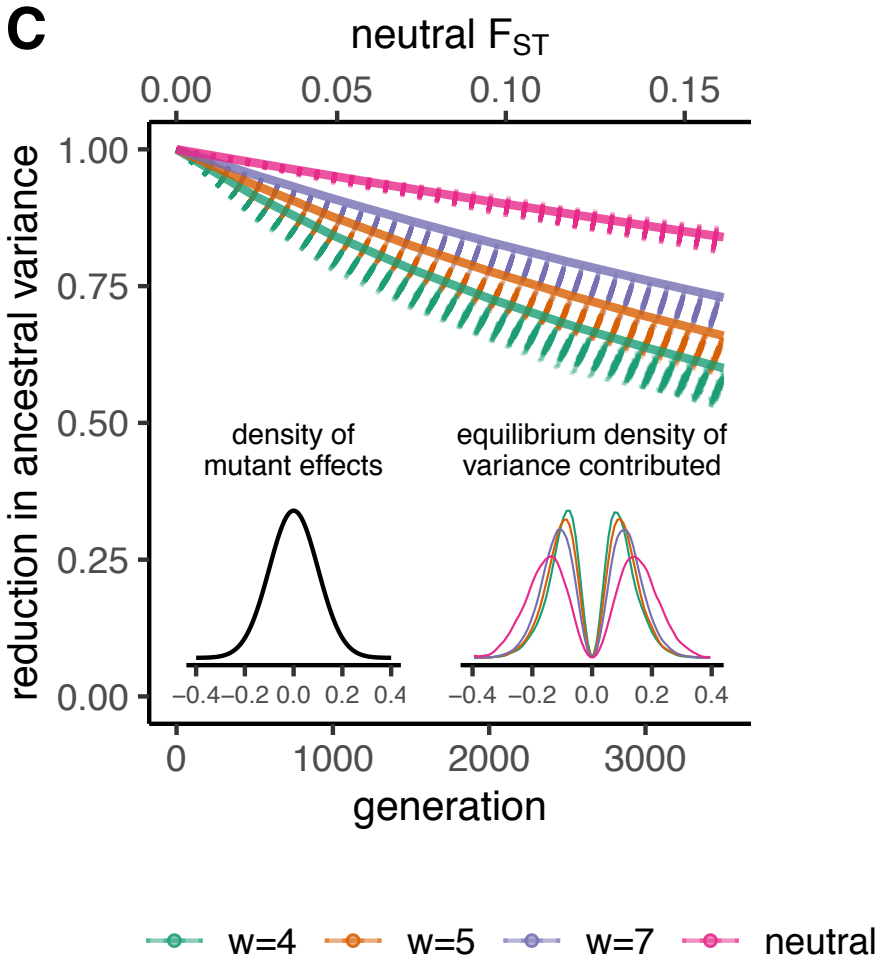
w=4 w=5 w=7 neutral

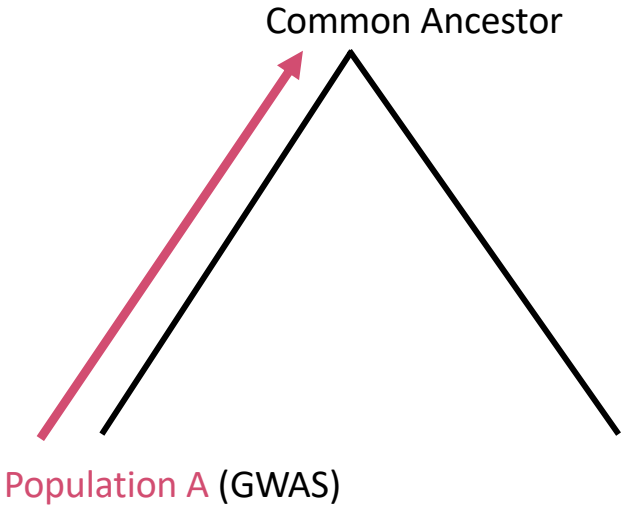

B
reduction in heterozygosity after 1500 generations
1.00
0.75
0.50
0.25
0.00
0.0
0.1
0.2
0.3
effect size
C
neutral F_{ST}
0.00
0.05
0.10
0.15
reduction in ancestral variance
1.00
0.75
0.50
0.25
0.00
density of mutant effects
equilibrium density of variance contributed
−0.4 −0.2 0.0 0.2 0.4
−0.4 −0.2 0.0 0.2 0.4
0
1000
2000
3000
generation
width of fitness peak
0.3
w=4
w=5
w=7
neutral

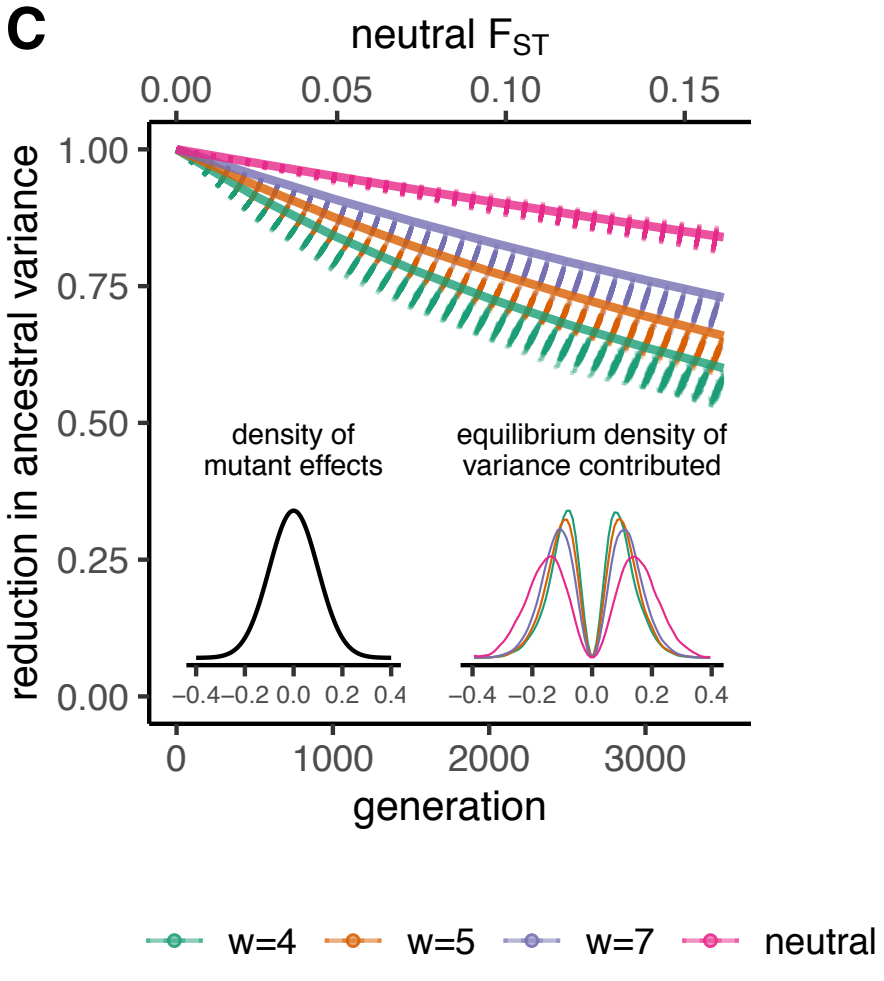


Prediction accuracy (r^2): proportion of variance in a population explained by PGS SNPs

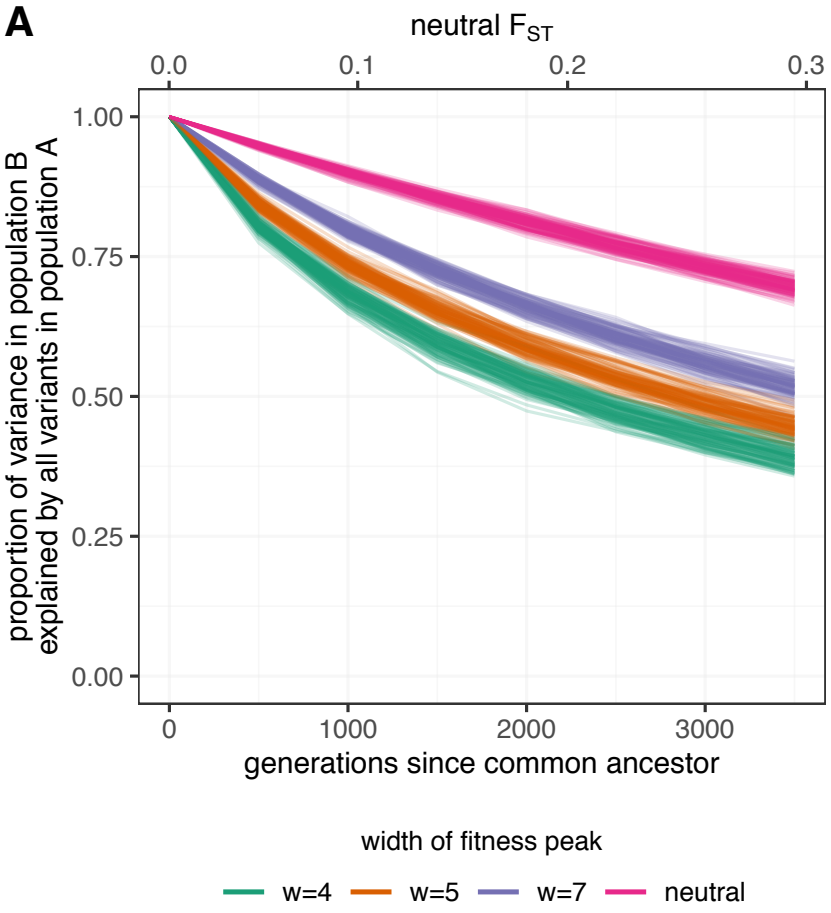
$$\frac{r^2_{PGS}}{r^2_{GV}} = \frac{V_{PGS}}{V_A} = \frac{\sum_{l \in \mathrm{PGS}} \alpha_l^2 p_l (1-p_l)}{\sum_{l \in \mathrm{All}} \alpha_l^2 p_l (1-p_l)}$$

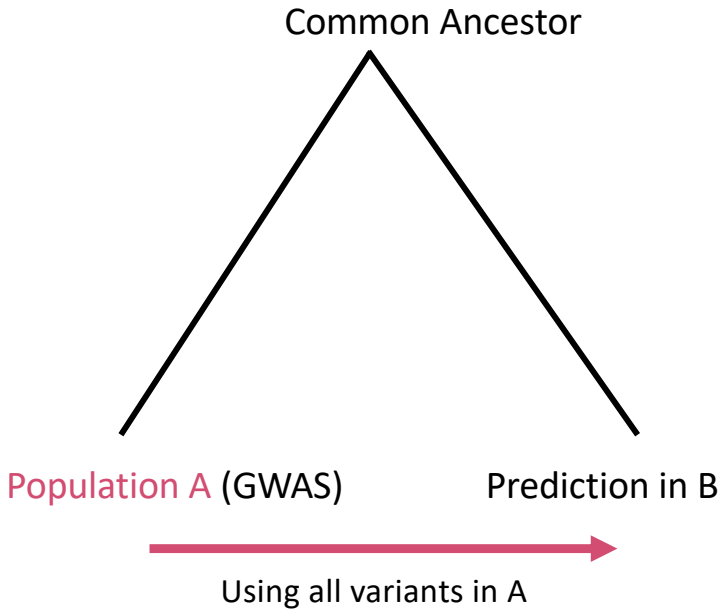


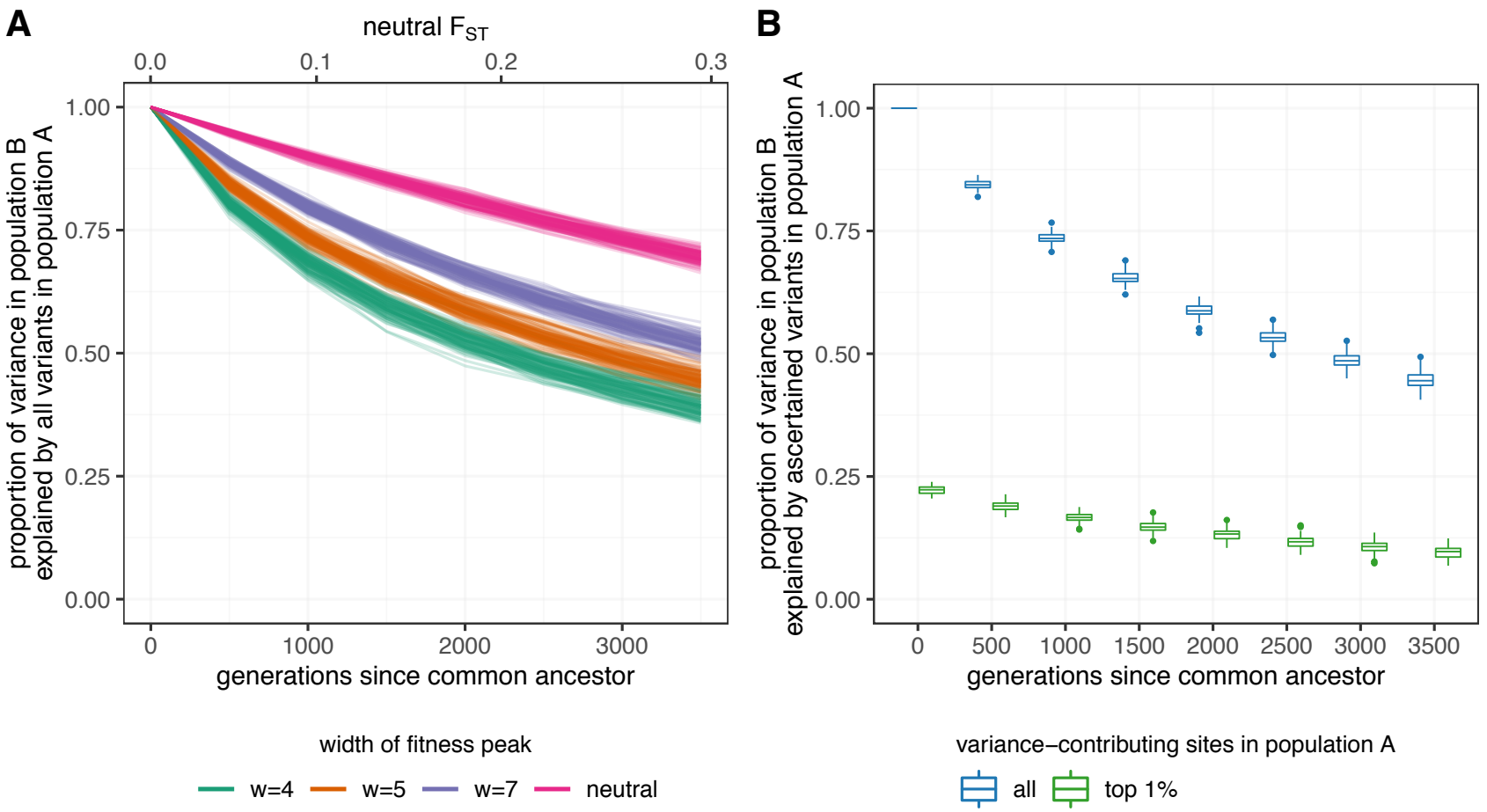




$$\frac{r^2_{PGS}}{r^2_{GV}} = \frac{V_{PGS}}{V_A} = \frac{\sum_{l \in \mathrm{PGS}} \alpha_l^2 p_l (1-p_l)}{\sum_{l \in \mathrm{All}} \alpha_l^2 p_l (1-p_l)}$$


A
neutral F_{ST}
0.0
0.1
0.2
0.3
1.00
0.75
0.50
0.25
0.00
proportion of variance in population B explained by all variants in population A
0
1000
2000
3000
generations since common ancestor
width of fitness peak
w=4
w=5
w=7
neutral


Common Ancestor
Population A (GWAS)
Prediction in B
Using all variants in A


A
neutral F_{ST}
0.0
0.1
0.2
0.3
proportion of variance in population B explained by all variants in population A
1.00
0.75
0.50
0.25
0.00
0
1000
2000
3000
generations since common ancestor
width of fitness peak
w=4
w=5
w=7
neutral
B
proportion of variance in population B explained by ascertained variants in population A
1.00
0.75
0.50
0.25
0.00
0
500
1000
1500
2000
2500
3000
3500
generations since common ancestor
variance−contributing sites in population A
all
top 1%

Stabilizing selection can drive more rapid turn over in GWAS variants lowering portability

Population differences in mean additive genetic values and mean polygenic scores

Additive genetic value of an individual

$$Z = \alpha_1 g_1 + \alpha_2 g_2 + \cdots + \alpha_L g_L$$

Mean additive genetic value of population

$$\bar{Z} = 2\alpha_1 p_1 + 2\alpha_2 p_2 + \cdots + 2\alpha_L p_L$$

The Apportionment of Human Diversity

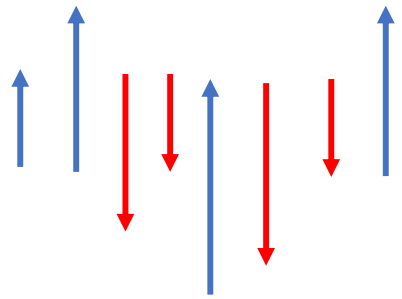
R. C. LEWONTIN

Committee on Evolutionary Biology, University of Chicago, Chicago, Illinois

1972

Estimated effect sizes: $\alpha_1, \alpha_2, \alpha_3 \ldots$
Genotypes $g_1, g_2, g_3 \ldots$
Population freq. $p_1, p_2, p_3 \ldots$

For a neutrally evolving phenotype $\mathbb{E}\left[(\bar{Z}_A - \bar{Z}_B)^2\right]$



Population differences in mean additive genetic values and mean polygenic scores

Additive genetic value of an individual

$$Z = \alpha_1 g_1 + \alpha_2 g_2 + \cdots + \alpha_L g_L$$

Mean additive genetic value of population

$$\bar{Z} = 2\alpha_1 p_1 + 2\alpha_2 p_2 + \cdots + 2\alpha_L p_L$$

The Apportionment of Human Diversity

R. C. LEWONTIN 1972

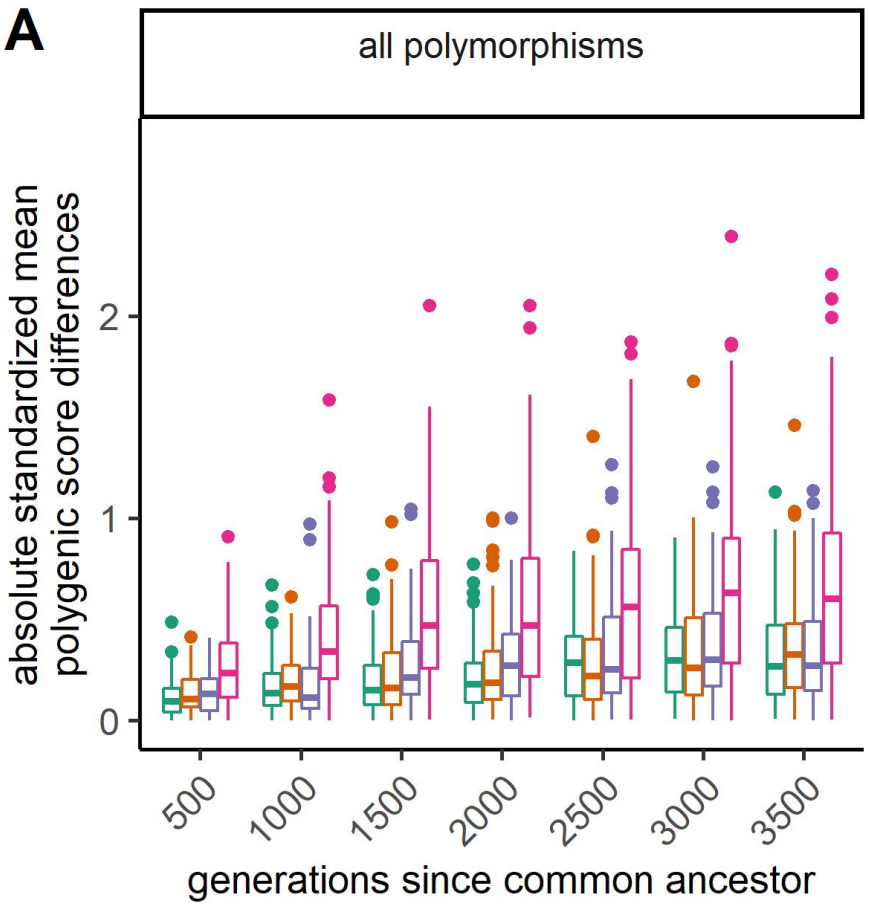
Committee on Evolutionary Biology, University of Chicago, Chicago, Illinois

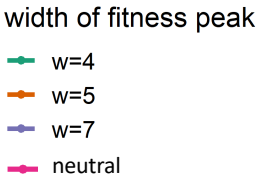
Estimated effect sizes: $\alpha_1, \alpha_2, \alpha_3 \ldots$
Genotypes $g_1, g_2, g_3 \ldots$
Population freq. $p_1, p_2, p_3 \ldots$

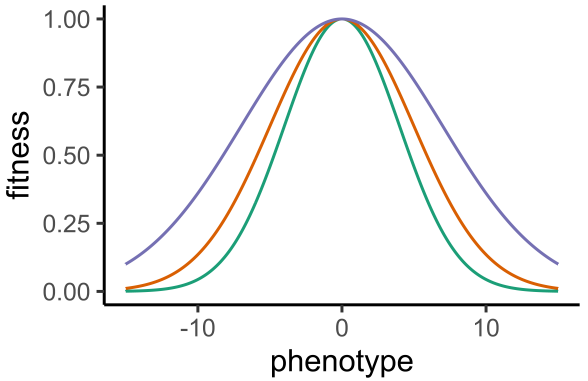
For a neutrally evolving phenotype $\mathbb{E}\left[(\bar{Z}_A - \bar{Z}_B)^2\right] = 4V_a F_{ST}$

Stabilizing selection is expected to reduce this.

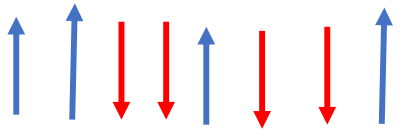
We'll study this using absolute standardized difference $\frac{|\bar{Z}_A - \bar{Z}_B|}{2\sqrt{V_a F_{ST}}}$

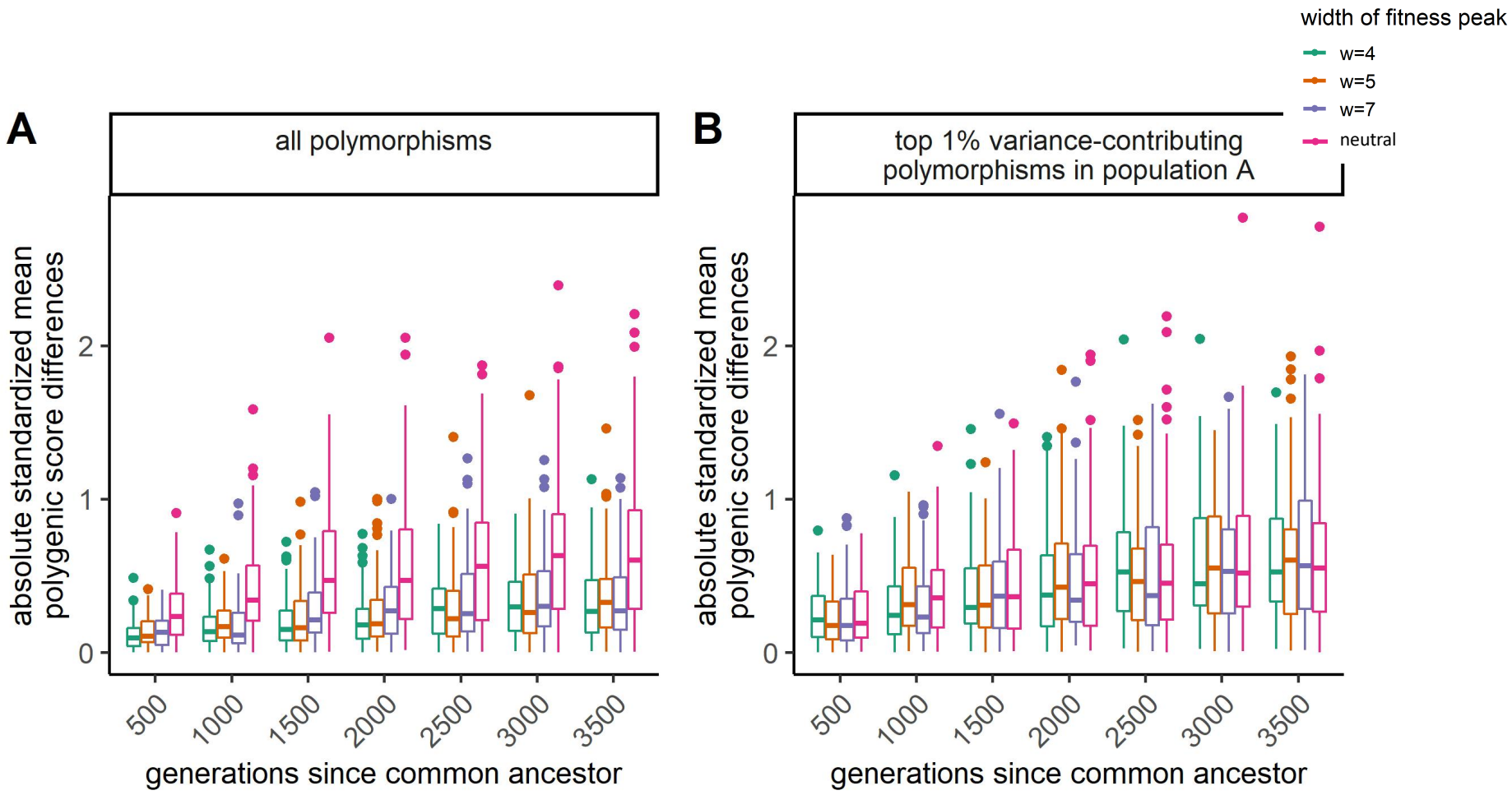


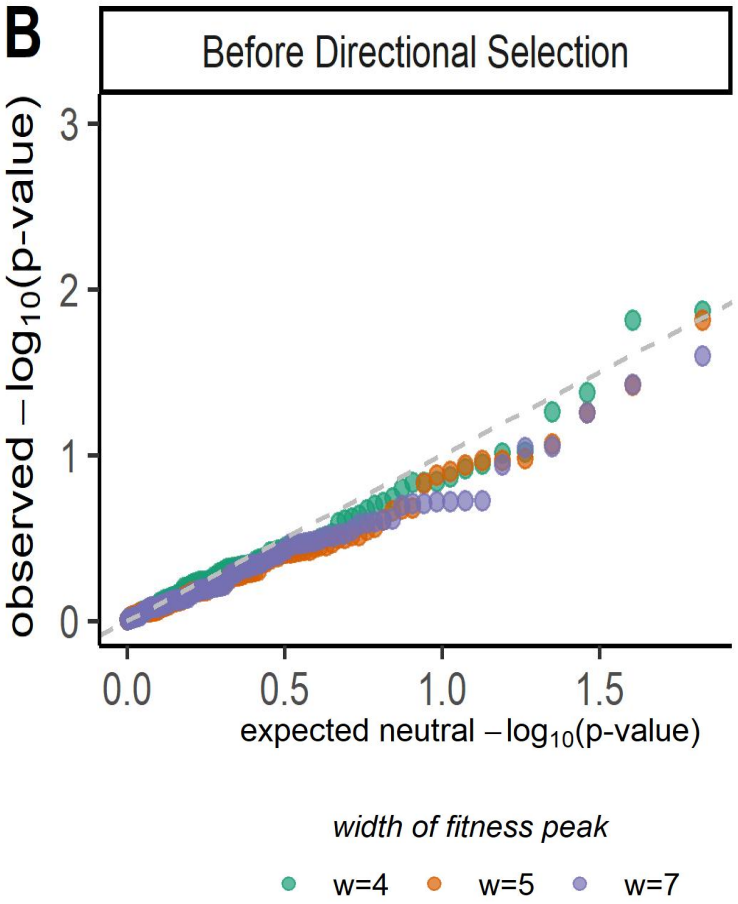


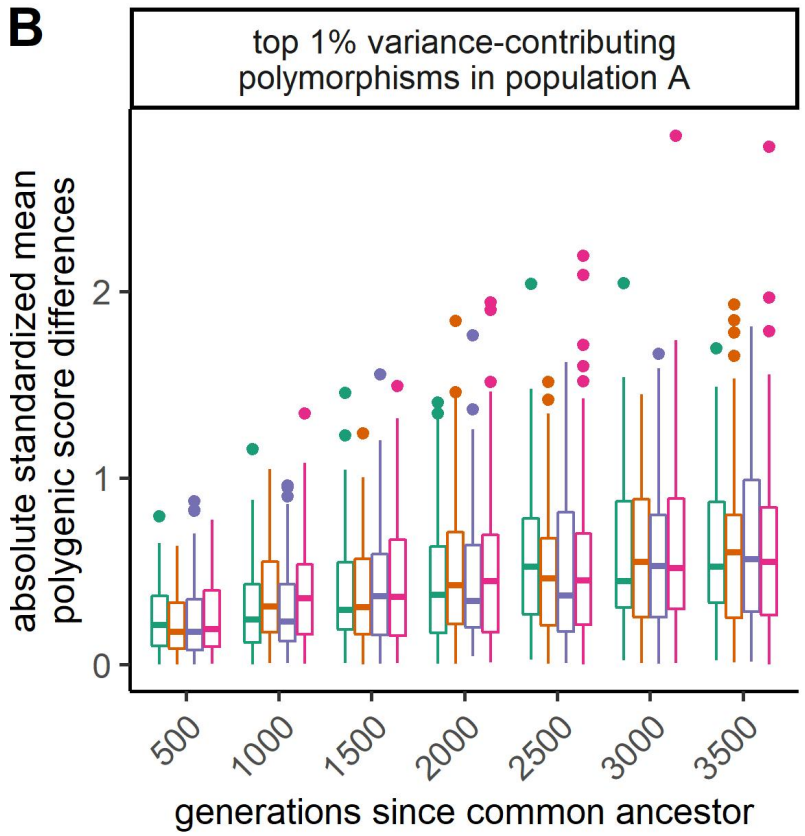


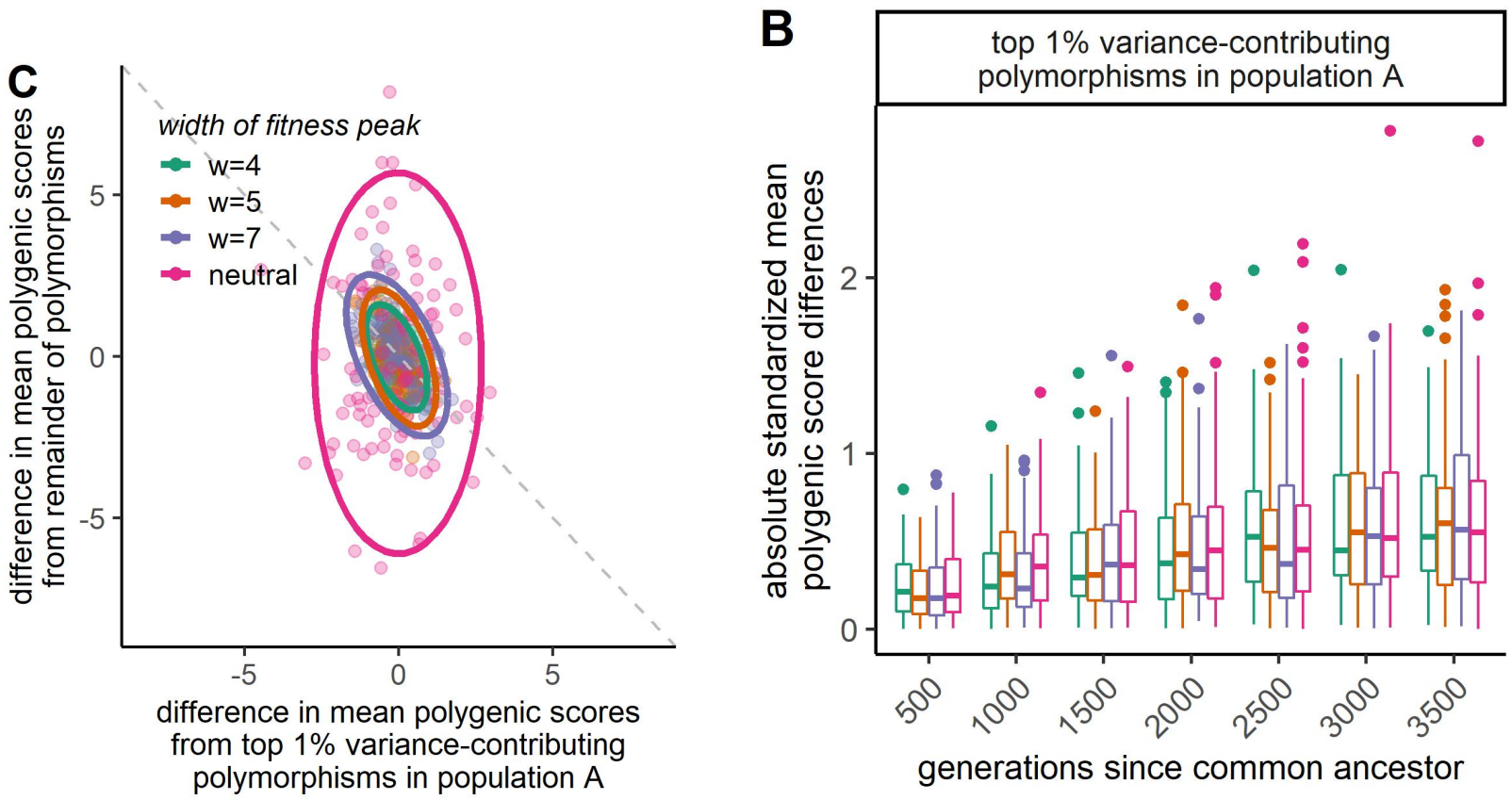
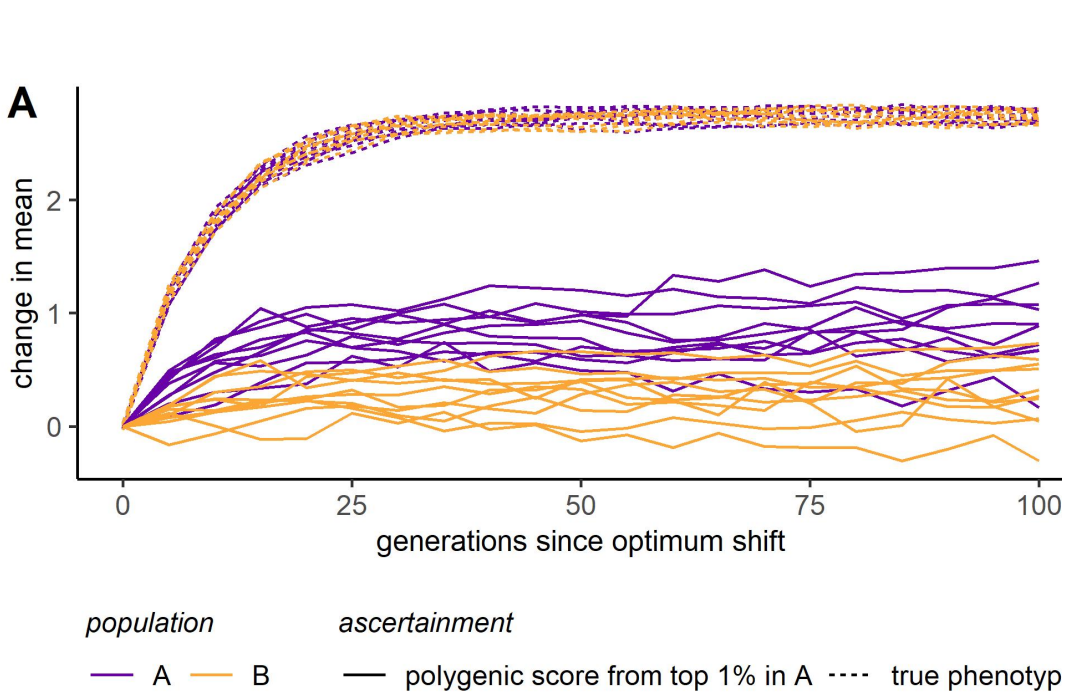
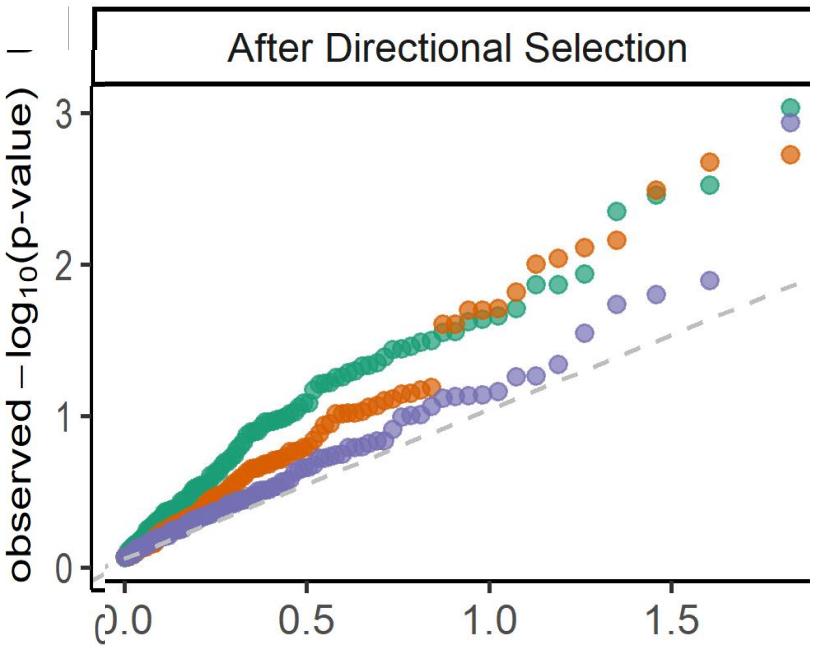
Why does stabilizing selection reduce the standardized mean difference of additive genetic values




A
all polymorphisms
B
top 1% variance-contributing
polymorphisms in population A
absolute standardized mean
polygenic score differences
0
1
2
500
1000
1500
2000
2500
3000
3500
generations since common ancestor
width of fitness peak
w=4
w=5
w=7
neutral


B
Before Directional Selection
observed −log10(p-value)
0
1
2
3
0.0
0.5
1.0
1.5
expected neutral −log10(p-value)
width of fitness peak
w=4
w=5
w=7


B
top 1% variance-contributing
polymorphisms in population A
absolute standardized mean
polygenic score differences
0
1
2
500
1000
1500
2000
2500
3000
3500
generations since common ancestor


C
width of fitness peak
w=4
w=5
w=7
neutral
difference in mean polygenic scores from remainder of polymorphisms
difference in mean polygenic scores from top 1% variance-contributing polymorphisms in population A
5
0
-5
-5
0
5
B
top 1% variance-contributing polymorphisms in population A
absolute standardized mean polygenic score differences
0
1
2
500
1000
1500
2000
2500
3000
3500
generations since common ancestor

False signals of differentiation under parallel directional selection