



**NORMENT**

Norwegian Centre for  
Mental Disorders Research

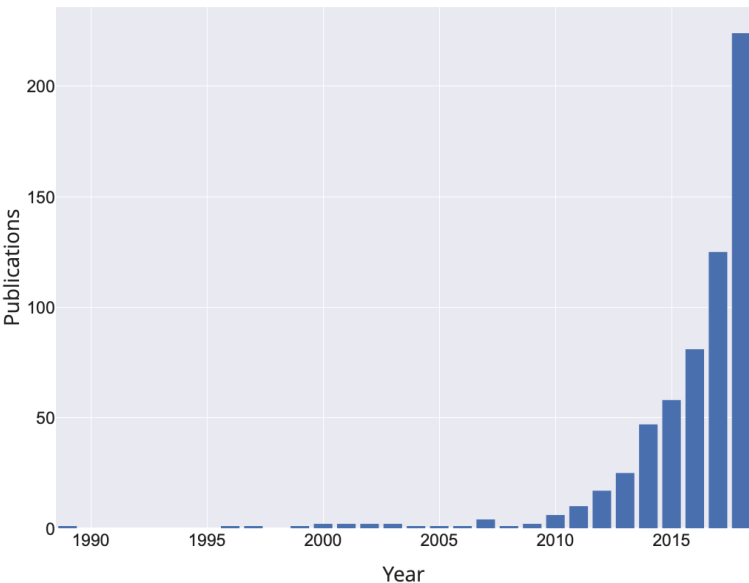


# Polygenic risk scores

Introduction and self-help guide

# Examples

*Papers tagged polygenic risk scores, per year, listed on PubMed*



Molecular Psychiatry  
https://doi.org/10.1038/s41380-018-0030-8

## ARTICLE

### Use of an Alzheimer's disease polygenic risk score to identify mild cognitive impairment in adults in their 50s

Mark W. Logue<sup>1,2,3</sup> · Matthew S. Panizzon<sup>4,5</sup> · Jeremy A. Elman<sup>6</sup> · Nathan A. Gillespie<sup>6</sup> · Sean N. Hatton<sup>4,5</sup> · Daniel E. Gustavson<sup>4,5</sup> · Ole A. Andreassen<sup>7,8</sup> · Anders M. Dale<sup>4,5,9,10</sup> · Carol E. Franz<sup>4,5</sup> · Michael J. Lyons<sup>11</sup> · Michael C. Neale<sup>6</sup> · Charles A. DeLisi<sup>12</sup> · Victor B. Jensen<sup>13</sup> · William C. Kremen<sup>4,5,14</sup>

European Journal of Human Genetics (2018) 26:1049–1059  
https://doi.org/10.1038/s41431-018-0134-2

## ARTICLE

### Effects of autozygosity and schizophrenia polygenic risk on cognitive and brain developmental trajectories

Aldo Córdova-Palomera<sup>1</sup> · Tobias Kaufmann<sup>1</sup> · Francesco Bettella<sup>1</sup> · Yunpeng Wang<sup>1</sup> · Nhat Trung Doan<sup>1</sup> · Dennis van der Meer<sup>1</sup> · Dag Alnæs<sup>1</sup> · Jaroslav Rokicki<sup>1,2</sup> · Torgeir Moberget<sup>1</sup> · Ida Elken Sørderby<sup>1</sup> · Ole A. Andreassen<sup>1</sup> · Lars T. Westlye<sup>1,2</sup>

## Acta Psychiatrica Scandinavica

Acta Psychiatr Scand 2014; 130: 311–317  
All rights reserved  
DOI: 10.1111/acps.12307

© 2014 John Wiley & Sons A/S. Published by John Wiley & Sons Ltd  
ACTA PSYCHIATRICA SCANDINAVICA

### Polygenic risk score and the psychosis continuum model

Tesli M, Espeseth T, Djurovic S, Andreassen O. Polygenic risk score (PGRS) to subcategories along the psychosis continuum model.

**Objective:** Schizophrenia and bipolar disorder are polygenic disorders with overlapping genetic risk. We investigated whether a polygenic risk score (PGRS) could be used to subcategories along the psychosis continuum model.



Contents lists available at ScienceDirect

Journal of Affective Disorders

journal homepage: [www.elsevier.com/locate/jad](http://www.elsevier.com/locate/jad)

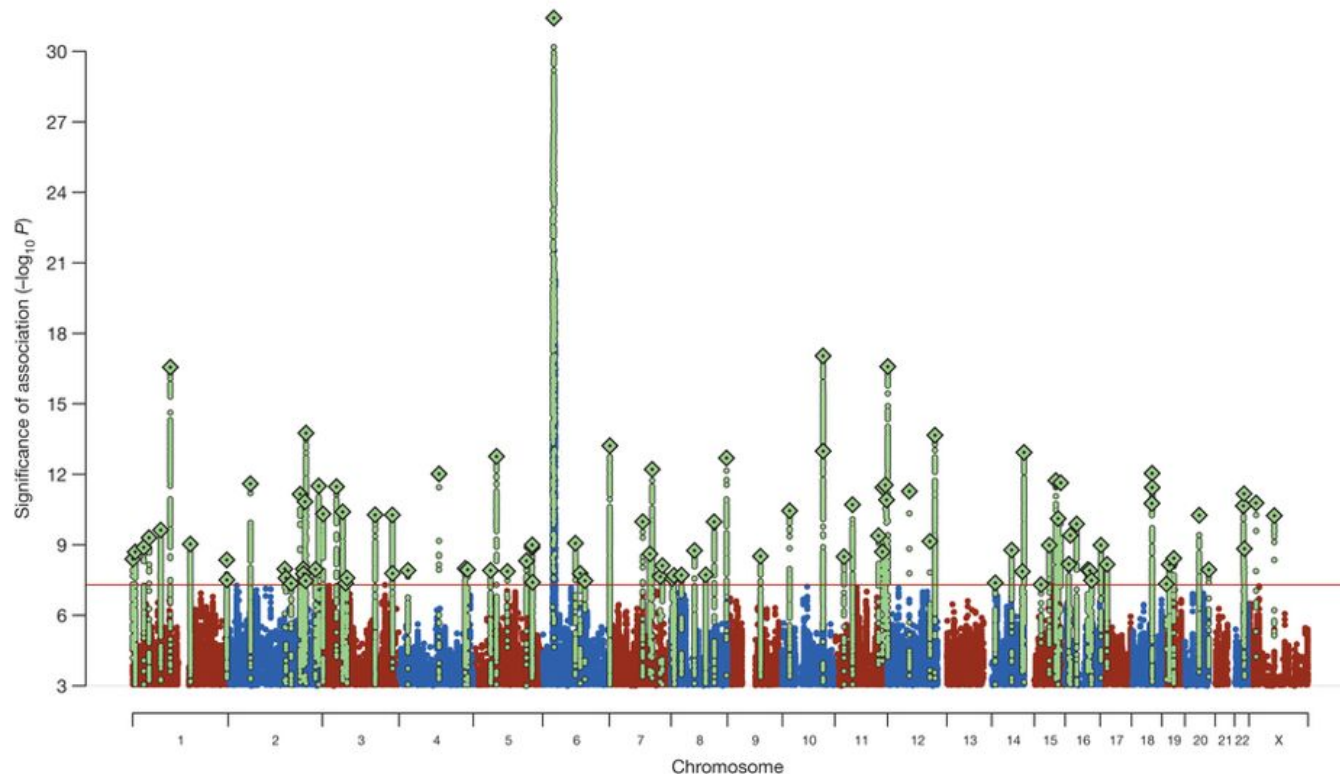
## Research report

### Polygenic risk scores in bipolar disorder subgroups

Sofie Ragnhild Aminoff<sup>a,b,\*</sup>, Martin Tesli<sup>b</sup>, Francesco Bettella<sup>b</sup>, Monica Aas<sup>b</sup>, Trine Vik Lagerberg<sup>b</sup>, Srdjan Djurovic<sup>b</sup>, Ole A. Andreassen<sup>b</sup>, Ingrid Melle<sup>b</sup>

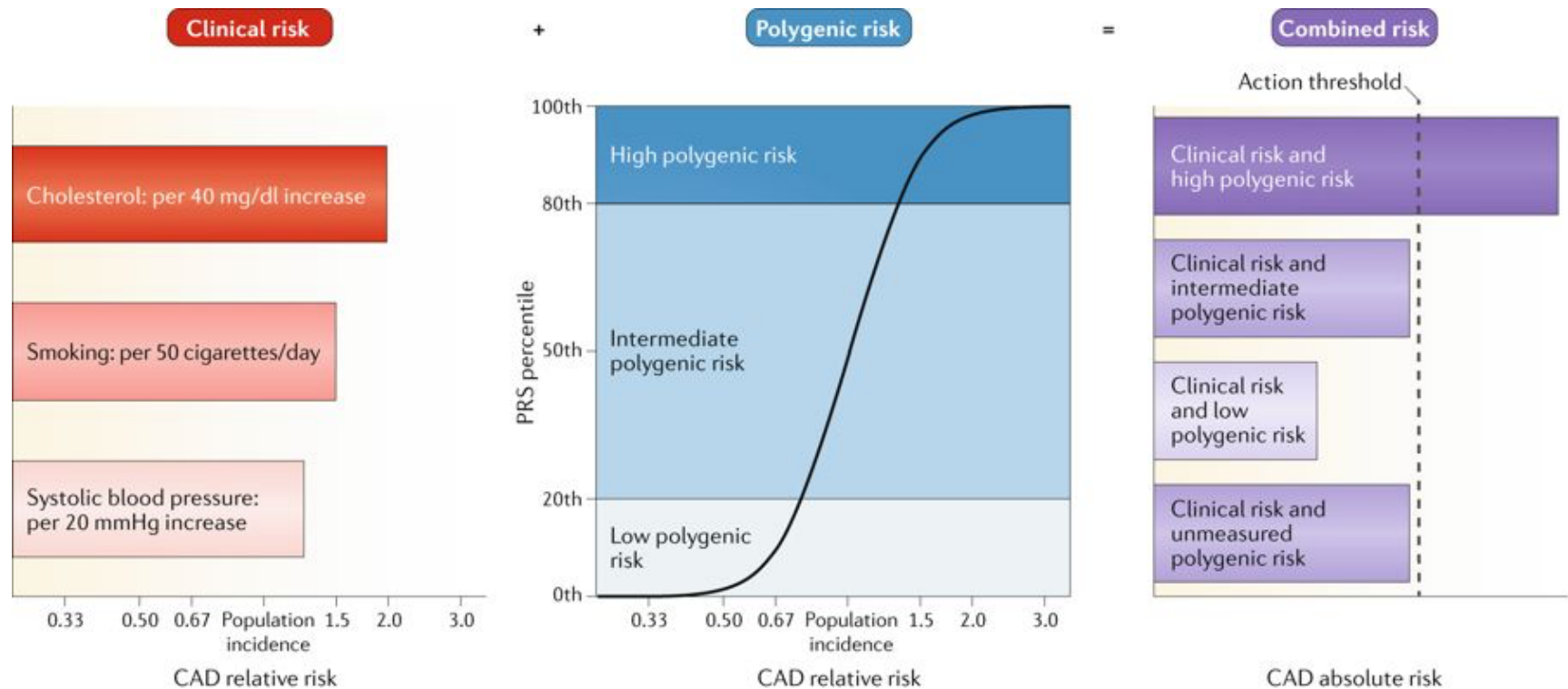
# Motivation

- Most phenotypes are polygenic, with single variants have very small effect
- The predictive power of a single variant is negligible
- Use instead the cumulative effect



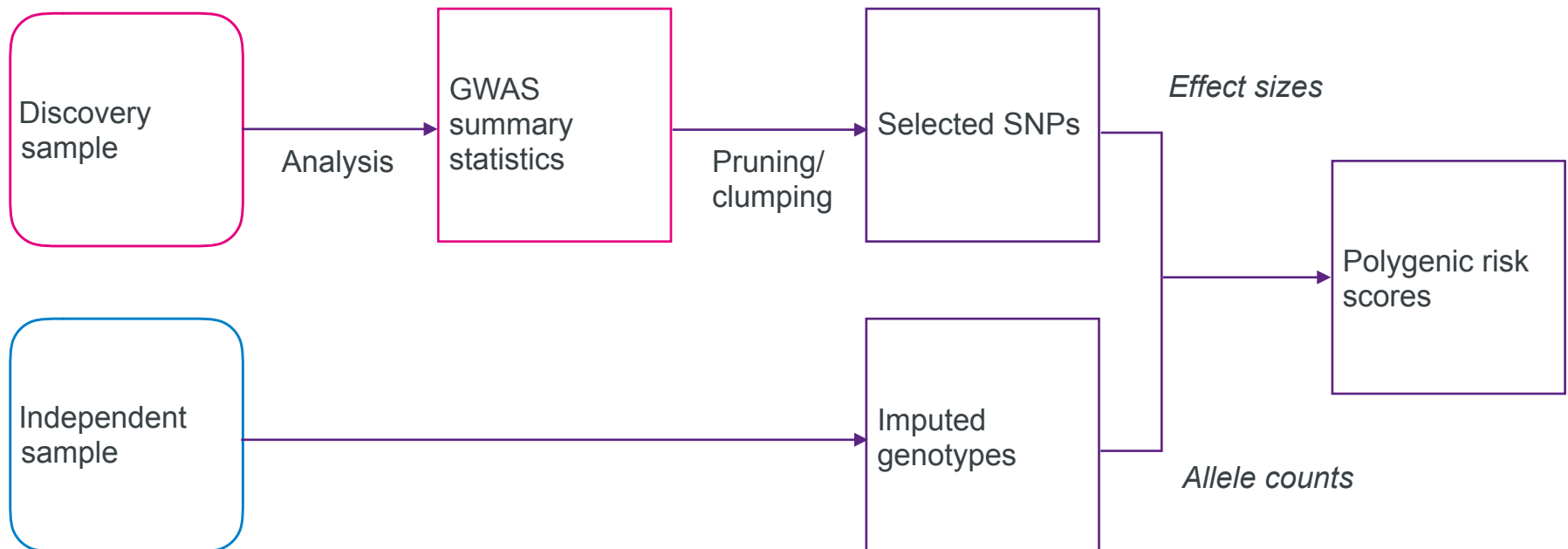
# Clinical use

*PGS-informed...* interventions – disease screening – life planning



# Pipeline

- Scores for a given phenotype are computed using SNP effect sizes from a relevant (meta-) GWAS



# Definition

An individual's polygenic risk score  $Y$  is defined as

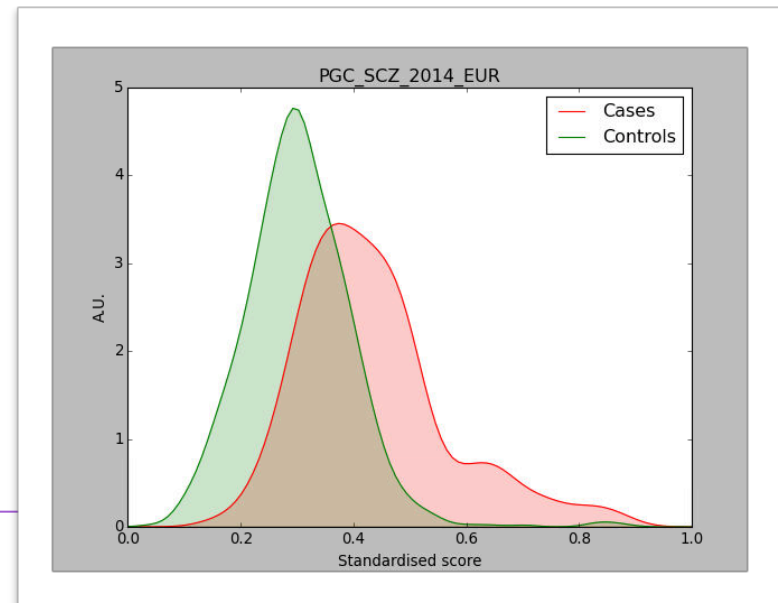
$$Y_{\text{PGS}} = \sum_i x_i \cdot \beta_i$$

$x_i$  : Allele count  $\{0,1,2\}$   
Measured by genotyping

$\beta_i$  : Effect size (log OR)  
Measured by GWAS

summing over all relevant SNPs  $i$

The *absolute* value of the score is not so informative, should be seen relative to the distribution of e.g. cases and controls

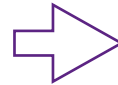


# Example

From GWAS, select SNPs below  $p$ -value threshold (e.g. 0.05)

SNP	A1	A2	$p$	$\beta$
rs915677	A	G	0.0217	-0.582
rs131564	C	G	0.0408	+0.666
rs4010550	G	A	0.0298	+0.674
rs11089263	A	C	0.0092	-1.308
rs11089264	A	G	0.0019	-1.266
rs2154615	T	C	0.0018	+1.276
rs5993628	A	G	0.0014	-1.098
rs2845362	C	G	0.0352	+0.399
...			...	...

**SNP list**



Remove correlated SNPs, keep only one representative SNP per LD block

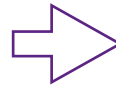
SNP	A1	A2	$p$	$\beta$
rs915677	A	G	0.0217	-0.582
rs2154615	T	C	0.0018	+1.276
rs2845362	C	G	0.0352	+0.399

**SNP loci**

# Example

## *SNP loci*

SNP	A1	A2	<i>p</i>	$\beta$
rs915677	A	G	0.0217	-0.582
rs2154615	T	C	0.0018	+1.276
rs2845362	C	G	0.0352	+0.399



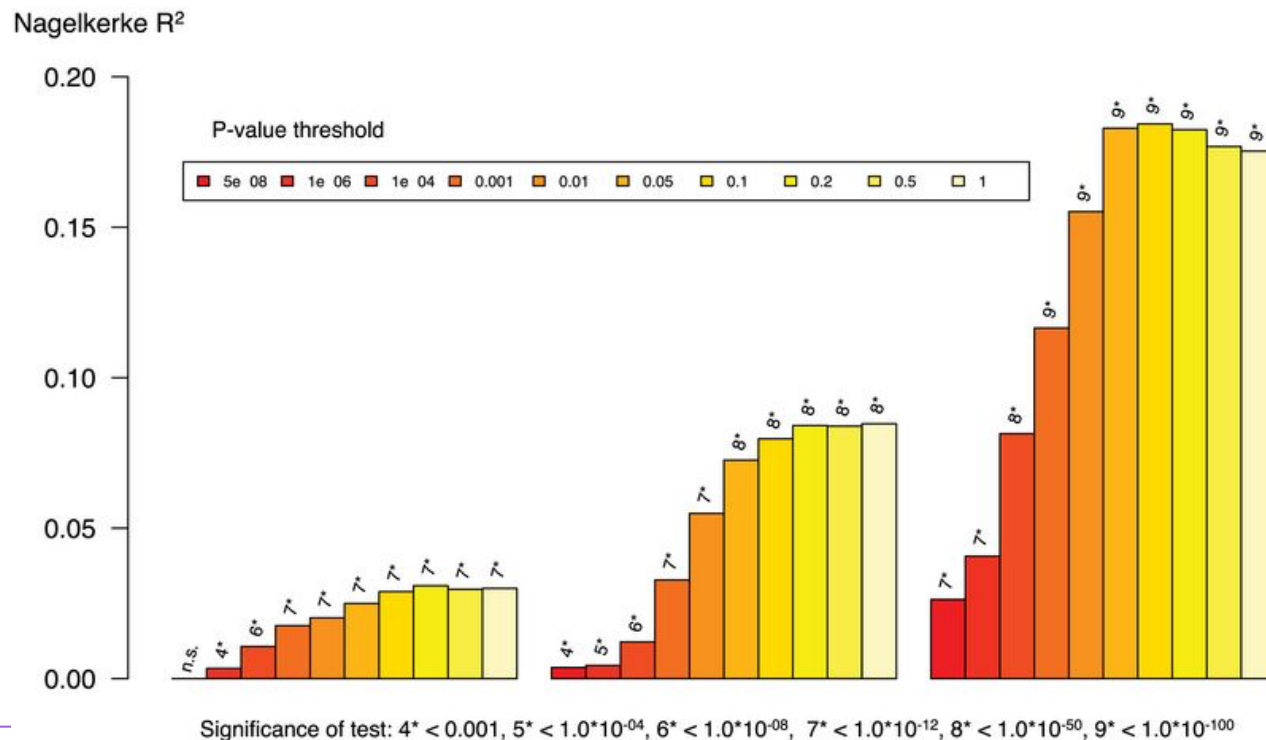
## *Kari Nordmann's genotype*

SNP	Genotype	$\beta$	<i>risk</i>
rs915677	AA	-0.582	-1.164
rs2154615	TC	+1.276	+1.276
rs2845362	GG	+0.399	0.0
PGS			+0.112



# Which SNPs to include?

- Optimal  $p$ -value threshold may differ between phenotypes
- Strict selection may remove SNPs with real effect, loose selection may include noise
  - Relatively loose selection ( $p < 0.1$ ) generally a good choice



# Technical walk-through

Following steps are required in practice:

0. Imputation
1. Quality control – both summary stats and genotypes  
*Remove poorly imputed variants*
2. Select lead SNPs for each LD block (clumping)  
*Take the most significant SNP (lowest p-value) to represent the block*
3. Ensure consistency between GWAS and genotypes  
*Make sure effects go in the right direction*
4. Compute scores

Tools required:

`plink`, `awk`, `bash`

Other PGS-specific software exists, e.g. `PRSice`

# Quality control

Typical requirements for summary statistics:

- Remove insertions or deletions
- Remove ambiguous SNPs (i.e. AT/GC pairs)
- Remove implausible  $p$ -values ( $= 0$ ) or effect sizes ( $OR > 10$ )
- Require imputation quality (INFO)  $> 0.8$
- Remove SNPs with low sample size

Typical requirements for imputed genotypes:

- Require minor allele frequency (MAF)  $> 1\%$
- Require SNP call rate  $> 0.9$
- Require imputation  $r^2 > 0.8$

rs62513865	C	T
rs79643588	G	A
rs17396518	T	G
rs983166	A	C
rs28842593	T	C
rs7014597	G	C
rs3134156	T	C
rs6980591	A	C
rs72670434	A	T
rs10955343	C	T
rs74749112	G	A
rs4734443	C	T
rs2436965	A	G
rs427145	A	G
rs7838119	G	A
rs4448250	T	C
rs2441902	G	A
rs34397009	C	G
rs11986967	T	C
rs9297348	T	C
rs16873161	C	T
rs56999386	T	C
rs402658	G	A
rs7008054	T	C
rs2154636	C	T
rs7010742	A	C
rs148863896	A	G
rs9297415	C	T
rs1786334	C	T
rs7002798	A	G

# Clumping

<https://www.cog-genomics.org/plink/1.9/>

- If multiple significant SNPs the same region, LD should be taken into account
- Requires a reference data set for LD structure (eg. 1000 Genomes)
- Set up PLINK on TSD:

```
module load plink
```

- Run:

```
plink --bfile [reference] --clump [input]
```

- Outputs a text file (plink.clumped) with one lead SNP per line, along with a list of SNPs in same LD block
- Relevant optional parameters:

```
--clump-field PVAL --clump-kb 250 --clump-p1 1.0  
--clump-p2 1.0 --clump-r2 0.25 --clump-verbose
```

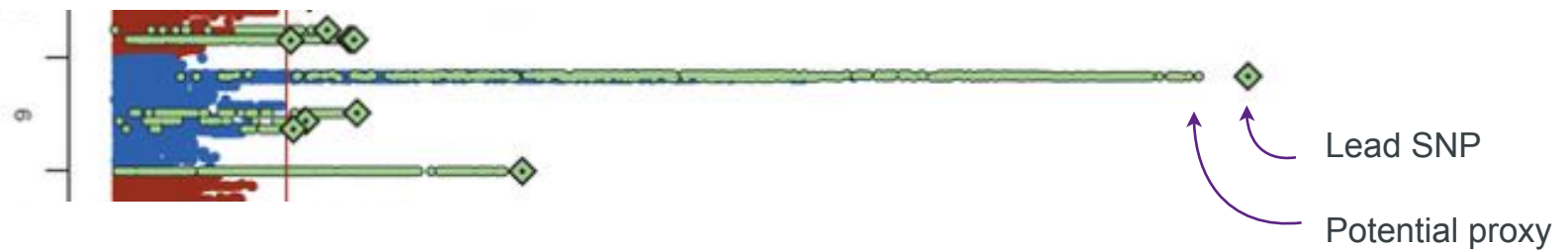
# Consistency checks

## *Strand flips:*

- If the summary statistics and the genotype data have opposite definitions of the alternate allele, one of them should be flipped

## *Proxy SNPs:*

- The available SNPs in the summary stats and the genotype data are often not completely overlapping
- If the lead SNP from the clumping step is not present in the genotypes, can select an other SNP in high LD with the lead SNP
  - Otherwise the effect from this LD block would be lost



# Compute scores

<https://www.cog-genomics.org/plink/1.9/>

- PLINK computes scores according to the additive model
- Requires a text file with significance thresholds for which SNPs to include, in the format

Output name	Lower p-value threshold	Upper p-value threshold
my_range_1	0.0	5.0e-8
my_range_2	0.0	0.01
...		

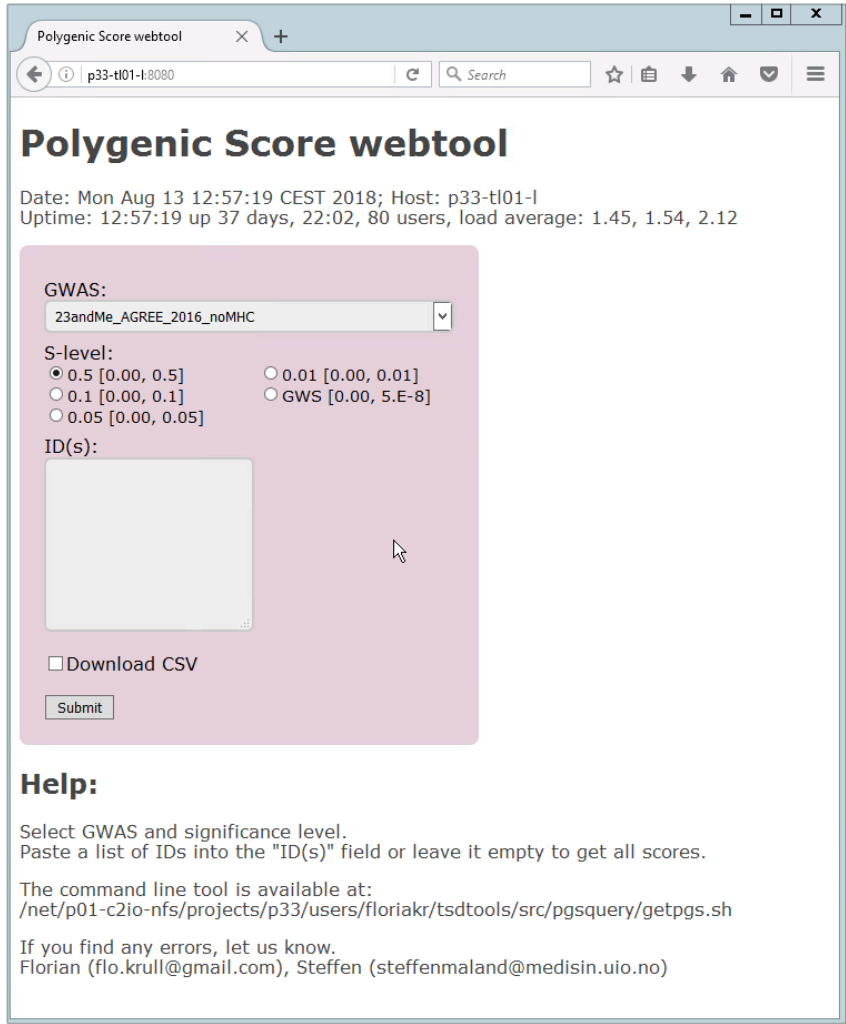
- Command slightly more complicated:

```
plink --dosage [dosage file] format=1 --fam [.fam file]
--score [cleaned sumstats file] [col modifiers]
--q-score-range [threshold file] [sumstat file]
[col modifiers]
```

- Outputs one file per threshold, with total score for each individual

# Downloads

- Scores for a range of phenotypes are available through a web interface on TSD
- Code is also available:  
`git clone http://p33-tl01-l:3000/steffma/polygenic-risk-score-pipeline`



The screenshot shows a web browser window titled "Polygenic Score webtool" with the address bar displaying "p33-tl01-l:8080". The page content includes a title "Polygenic Score webtool", a status bar with date, host, uptime, and load average, and a main form area. The form has a "GWAS:" dropdown menu set to "23andMe\_AGREE\_2016\_noMHC", an "S-level:" section with radio buttons for 0.5, 0.1, 0.05, 0.01, and GWS, an "ID(s):" text input field, a "Download CSV" checkbox, and a "Submit" button. Below the form is a "Help:" section with instructions on how to use the tool and contact information.

**Polygenic Score webtool**

Date: Mon Aug 13 12:57:19 CEST 2018; Host: p33-tl01-l  
Uptime: 12:57:19 up 37 days, 22:02, 80 users, load average: 1.45, 1.54, 2.12

**GWAS:**  
23andMe\_AGREE\_2016\_noMHC

**S-level:**  
☒ 0.5 [0.00, 0.5] ☐ 0.01 [0.00, 0.01]  
☐ 0.1 [0.00, 0.1] ☐ GWS [0.00, 5.E-8]  
☐ 0.05 [0.00, 0.05]

**ID(s):**

☐ Download CSV

**Help:**  
Select GWAS and significance level.  
Paste a list of IDs into the "ID(s)" field or leave it empty to get all scores.  
The command line tool is available at:  
/net/p01-c2io-nfs/projects/p33/users/floriakr/tsdtools/src/pgsquery/getpgs.sh  
If you find any errors, let us know.  
Florian (flo.krull@gmail.com), Steffen (steffenmaland@medisin.uio.no)

# Words of caution

- Estimation sample (GWAS) and prediction sample (for scores) must be independent
  - *Example:*  
Computing schizophrenia scores for TOP participants, using PGC data
    - **Not OK:** Ignoring the fact that TOP data is part PGC
    - **OK:** Excluding TOP from PGC data, then computing scores
- The quality of the scores is only as good as the GWAS they're based on
- If the GWAS is performed on a specific population group, scores only make sense for the same group
- Genetic principal components should be included if necessary
- Performance still too low to be useful on individual level



# Recent advancements

- Bayesian approach based on heritability and fraction of causal SNPs (LDpred):  
BJ. Vilhjálmsson et al: *Modeling Linkage Disequilibrium Increases Accuracy of Polygenic Risk Scores*. Am J Hum Genet. 97 (2015)
- Empirical Bayes approach:  
HC Soa, PC. Sham: *Improving polygenic risk prediction from summary statistics by an empirical Bayes approach*, Sci Rep. 7, 41262 (2017)
- Using boosted regression trees:  
G. Paré, S. Mao, WQ. Deng: *A machine-learning heuristic to improve gene score prediction of polygenic traits*. Scientific Reports 7, 12665 (2017)
- Including annotations:  
23andMe research team: *Modeling functional enrichment improves polygenic prediction accuracy in UK Biobank and 23andMe data sets*. <https://doi.org/10.1101/375337> (2018)



**NORMENT**

Norwegian Centre for  
Mental Disorders Research

Oslo University Hospital HF  
Division of Mental Health and Addiction  
Psychosis Research Unit/TOP  
Ullevål Hospital, building 49  
P.O. Box 4956 Nydalen  
N-0424 Oslo  
Norway