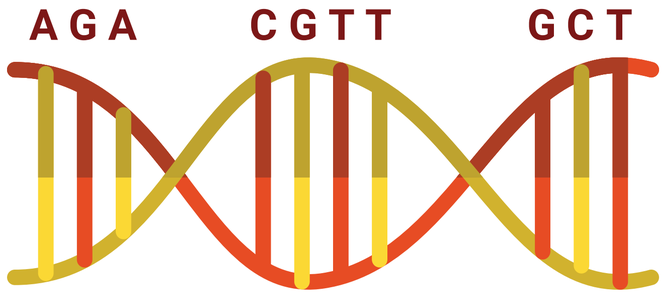
Private information leakage from polygenic risk scores

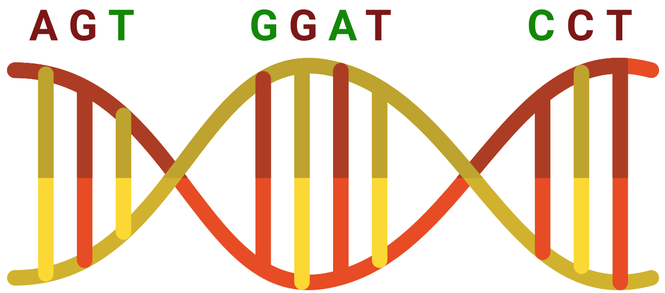
Kirill Nikitin*, Gamze Gürsoy†

*Columbia University & New York Genome Center, †University of Cambridge

RECOMB 2026

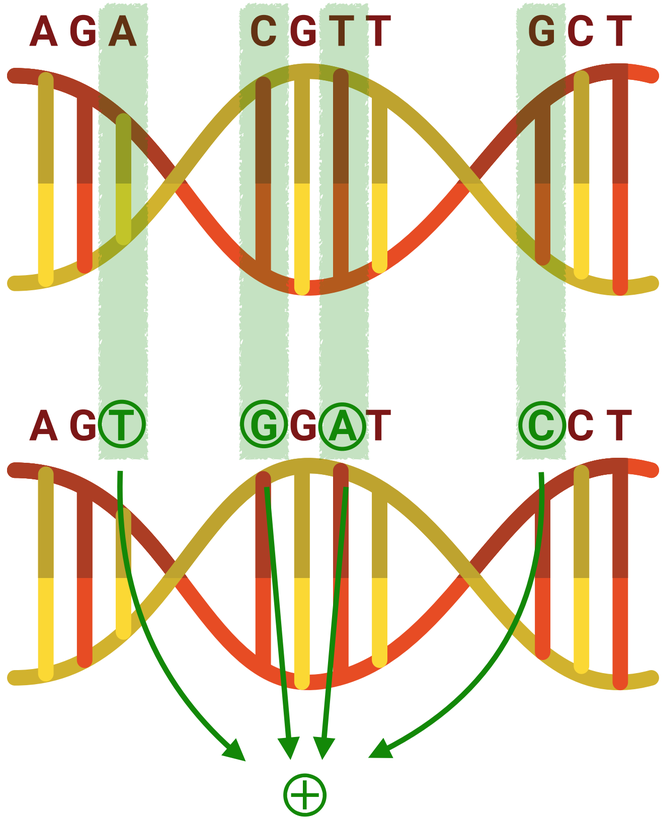
Polygenic Risk Score (PRS)



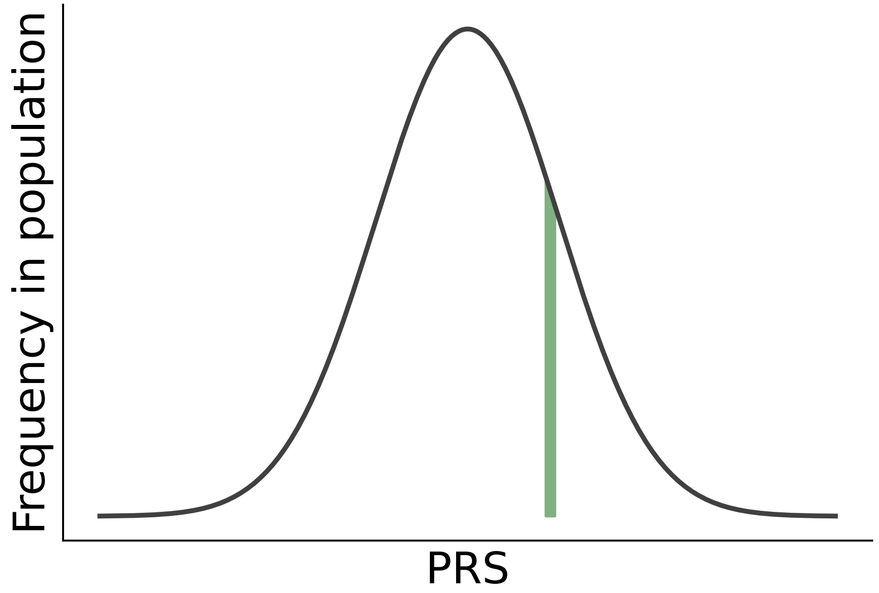


Complex traits: combined effect of multiple SNPs

Polygenic Risk Score (PRS)

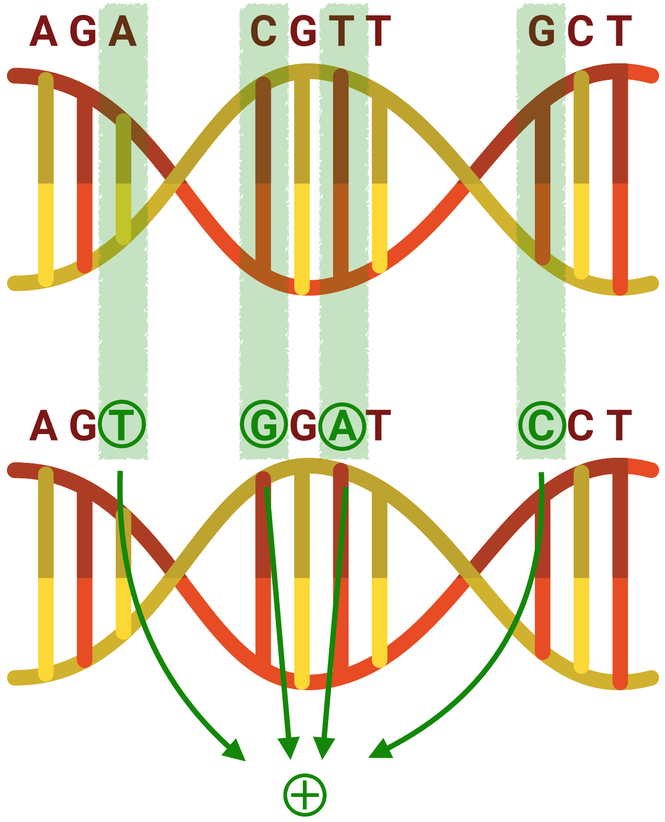


PRS = 3.5193571

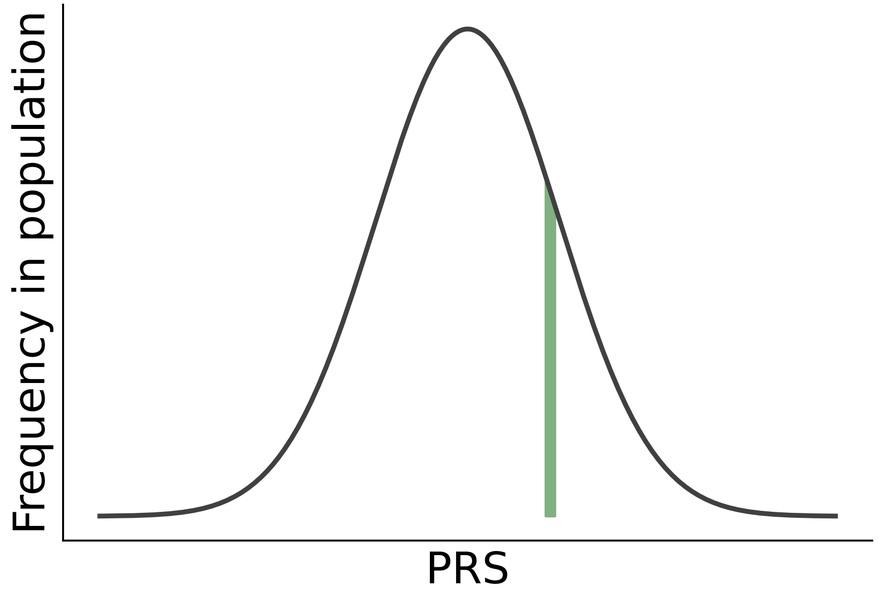


Complex traits: combined effect of multiple SNPs

Polygenic Risk Score (PRS)



PRS = 3.5193571



Carcinomas, lipid levels, diabetes, schizophrenia, heart diseases, …

Complex traits: combined effect of multiple SNPs

The promise of PRS

CLINICAL BREAKTHROUGHS

Deploying Polygenic Risk Scores in Primary Care Settings

By Melissa Rohman – Apr 17, 2024

Original Investigation FREE

January 30, 2019

Association of Polygenic Liabilities for Major Depression, Bipolar Disorder, and Schizophrenia With Risk for Depression in the Danish Population

Katherine L. Musliner, PhD[1,2]; Preben B. Mortensen, DrMedSc[1,2,3]; John J. McGrath, PhD[2,4,5]; et al

Author Affiliations | Article Information

JAMA Psychiatry. 2019;76(5):516-525. doi:10.1001/jamapsychiatry.2018.4166

ARTICLE

Polygenic Risk Scores for Prediction of Breast Cancer and Breast Cancer Subtypes

The promise of PRS

CLINICAL BREAKTHROUGHS

Deploying Polygenic Risk Scores in Primary Care Settings

By Melissa Rohman – Apr 17, 2024

Original Investigation FREE

January 30, 2019

Association of Polygenic Liabilities for Major Depression, Bipolar Disorder, and Schizophrenia With Risk for Depression in the Danish Population

Katherine L. Musliner, PhD[1,2]; Preben B. Mortensen, DrMedSc[1,2,3]; John J. McGrath, PhD[2,4,5]; et al

Author Affiliations | Article Information

JAMA Psychiatry. 2019;76(5):516-525. doi:10.1001/jamapsychiatry.2018.4166

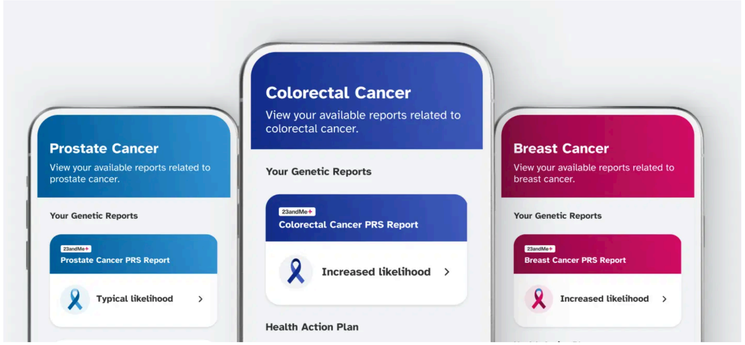
ARTICLE

Polygenic Risk Scores for Prediction of Breast Cancer and Breast Cancer Subtypes

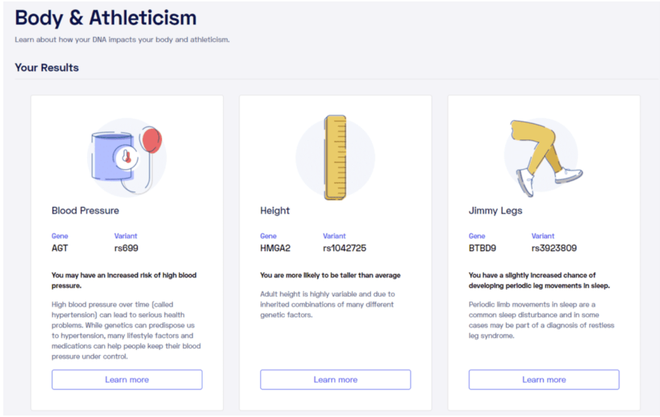
MAR 6, 2024 - HEALTH + TRAITS

23andMe Launches New Genetic Reports on Common Forms of Cancer

By **23andMe Blog Team** · 23andMe Blog



Nebula Genomics



Ethical and policy discussions around PRS

PERSPECTIVE

Potential corporate uses of polygenic indexes: Starting a conversation about the associated ethics and policy issues

Michelle N. Meyer,[1,*] Nicholas W. Papageorge,[2,*] Erik Parens,[3] Alan Regenberg,[4] Jeremy Sugarman,[4,5] and Kevin Thom[6]

Future implications of polygenic risk scores for life insurance underwriting

Tatiane Yanes ✉, Jane Tiller, Casey M. Haining, Courtney Wallingford, Margaret Otlowski, Louise Keogh, Aideen McInerney-Leo & Paul Lacaze

npj Genomic Medicine **9**, Article number: 25 (2024) | Cite this article

Ethical and policy discussions around PRS

PERSPECTIVE

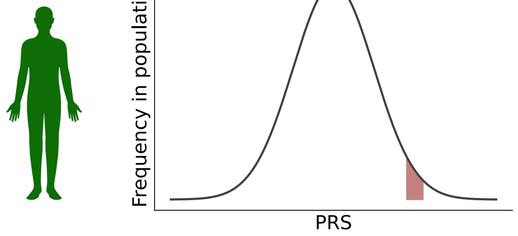
Potential corporate uses of polygenic indexes: Starting a conversation about the associated ethics and policy issues

Michelle N. Meyer,[1,*] Nicholas W. Papageorge,[2,*] Erik Parens,[3] Alan Regenberg,[4] Jeremy Sugarman,[4,5] and Kevin Thom[6]

Future implications of polygenic risk scores for life insurance underwriting

Tatiane Yanes ✉, Jane Tiller, Casey M. Haining, Courtney Wallingford, Margaret Otlowski, Louise Keogh, Aideen McInerney-Leo & Paul Lacaze

npj Genomic Medicine **9**, Article number: 25 (2024) | Cite this article



MetLife

AXA

PING AN

Ethical and policy discussions around PRS

PERSPECTIVE

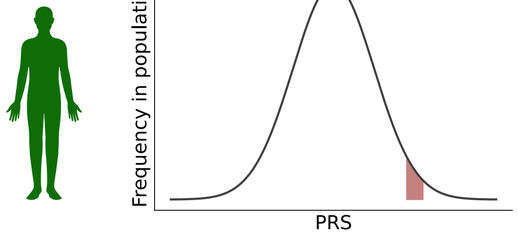
Potential corporate uses of polygenic indexes:
Starting a conversation about
the associated ethics and policy issues

Michelle N. Meyer,[1,*] Nicholas W. Papageorge,[2,*] Erik Parens,[3] Alan Regenberg,[4] Jeremy Sugarman,[4,5] and Kevin Thom[6]

Future implications of polygenic risk scores for life insurance underwriting

Tatiane Yanes, Jane Tiller, Casey M. Haining, Courtney Wallingford, Margaret Otlowski, Louise Keogh, Aideen McInerney-Leo & Paul Lacaze

npj Genomic Medicine **9**, Article number: 25 (2024) | Cite this article



MetLife

AXA

PING AN

- 🇺🇸 GINA does **not** cover life, disability or property insurances
- 🇦🇺 Protection does **not** extend to private health insurance
- 🇬🇧 Insurance providers **may** use existing genetic test results

Ethical and policy discussions around PRS

PERSPECTIVE

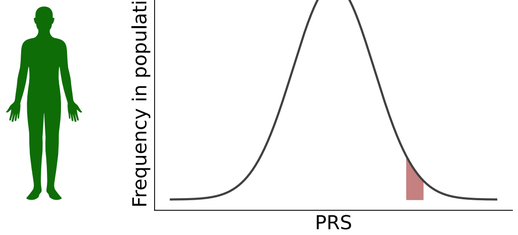
Potential corporate uses of polygenic indexes:
Starting a conversation about
the associated ethics and policy issues

Michelle N. Meyer,[1,*] Nicholas W. Papageorge,[2,*] Erik Parens,[3] Alan Regenberg,[4] Jeremy Sugarman,[4,5] and Kevin Thom[6]

Future implications of polygenic risk scores for life insurance underwriting

Tatiane Yanes, Jane Tiller, Casey M. Haining, Courtney Wallingford, Margaret Otlowski, Louise Keogh, Aideen McInerney-Leo & Paul Lacaze

npj Genomic Medicine **9**, Article number: 25 (2024) | Cite this article



MetLife

AXA

PING AN

- GINA does **not** cover life, disability or property insurances
- Protection does **not** extend to private health insurance
- Insurance providers **may** use existing genetic test results

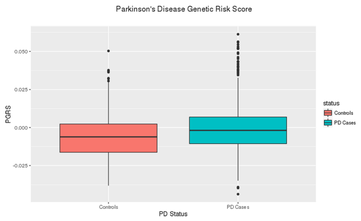
What are the *privacy implications* of PRSs?

Lack of PRS privacy guidance

Lack of PRS privacy guidance

Research publishing, clinical studies, …

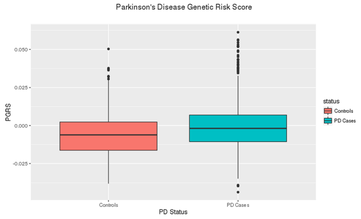
	A	B	C	D
1	**Cohort**	**Case control**	**Disease subtype**	**Standardized PRS**
2	clinical	1	Infiltrating	0.291159147
3	clinical	1	Infiltrating	0.706357435
4	clinical	1	Infiltrating	0.291159147
5	clinical	1	Infiltrating	-0.539237428
6	clinical	1	Infiltrating	0.291159147
7	clinical	1	Infiltrating	0.291159147
8	clinical	1	Infiltrating	1.951952297
9	clinical	1	Infiltrating	0.291159147
10	clinical	1	Infiltrating	1.53675401
11	clinical	1	Infiltrating	1.53675401
12	clinical	1	Infiltrating	-0.539237428
13	clinical	1	Infiltrating	0.706357435



Lack of PRS privacy guidance

Research publishing, clinical studies, …

	A	B	C	D
1	**Cohort**	**Case control**	**Disease subtype**	**Standardized PRS**
2	clinical	1	Infiltrating	0.291159147
3	clinical	1	Infiltrating	0.706357435
4	clinical	1	Infiltrating	0.291159147
5	clinical	1	Infiltrating	-0.539237428
6	clinical	1	Infiltrating	0.291159147
7	clinical	1	Infiltrating	0.291159147
8	clinical	1	Infiltrating	1.951952297
9	clinical	1	Infiltrating	0.291159147
10	clinical	1	Infiltrating	1.53675401
11	clinical	1	Infiltrating	1.53675401
12	clinical	1	Infiltrating	-0.539237428
13	clinical	1	Infiltrating	0.706357435



Online platforms

r/genomics • 4 yr. ago
Professional_Lead958

Help! Nebula results

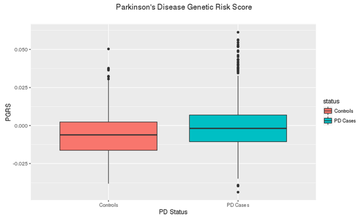
Did Nebula Genomics test and am freaking out — looking for a genetic counselor I can speak with.

by **Beatrice3** » Sun Dec 31, 2023 11:17 am

Lack of PRS privacy guidance

Research publishing, clinical studies, …

	A	B	C	D
1	Cohort	Case control	Disease subtype	Standardized PRS
2	clinical	1	Infiltrating	0.291159147
3	clinical	1	Infiltrating	0.706357435
4	clinical	1	Infiltrating	0.291159147
5	clinical	1	Infiltrating	-0.539237428
6	clinical	1	Infiltrating	0.291159147
7	clinical	1	Infiltrating	0.291159147
8	clinical	1	Infiltrating	1.951952297
9	clinical	1	Infiltrating	0.291159147
10	clinical	1	Infiltrating	1.53675401
11	clinical	1	Infiltrating	1.53675401
12	clinical	1	Infiltrating	-0.539237428
13	clinical	1	Infiltrating	0.706357435



Online platforms

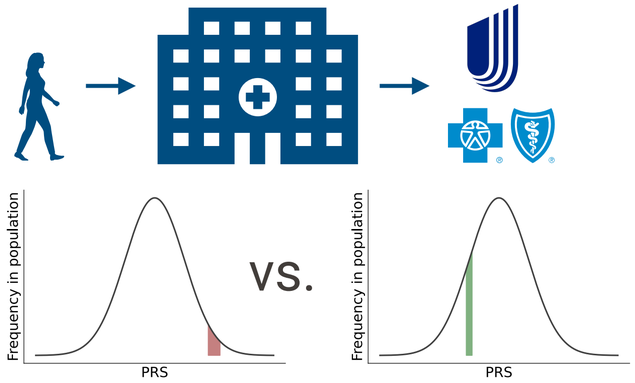
r/genomics • 4 yr. ago
Professional_Lead958

Help! Nebula results

Did Nebula Genomics test and am freaking out — looking for a genetic counselor I can speak with.

by **Beatrice3** » Sun Dec 31, 2023 11:17 am

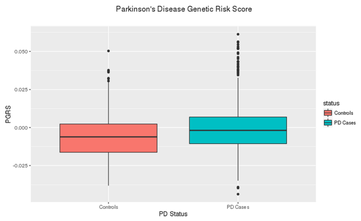
Doctor offices / insurance



Lack of PRS privacy guidance

Research publishing, clinical studies, …

	A	B	C	D
1	Cohort	Case control	Disease subtype	Standardized PRS
2	clinical	1	Infiltrating	0.291159147
3	clinical	1	Infiltrating	0.706357435
4	clinical	1	Infiltrating	0.291159147
5	clinical	1	Infiltrating	-0.539237428
6	clinical	1	Infiltrating	0.291159147
7	clinical	1	Infiltrating	0.291159147
8	clinical	1	Infiltrating	1.951952297
9	clinical	1	Infiltrating	0.291159147
10	clinical	1	Infiltrating	1.53675401
11	clinical	1	Infiltrating	1.53675401
12	clinical	1	Infiltrating	-0.539237428
13	clinical	1	Infiltrating	0.706357435



Online platforms

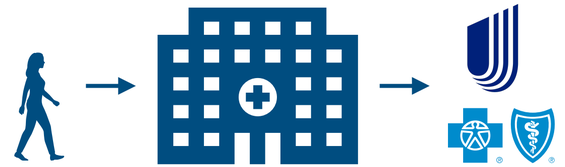
r/genomics • 4 yr. ago
Professional_Lead958

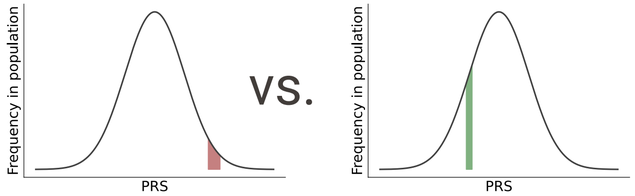
Help! Nebula results

Did Nebula Genomics test and am freaking out — looking for a genetic counselor I can speak with.

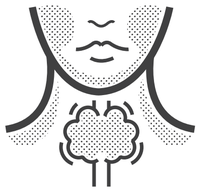
by **Beatrice3** » Sun Dec 31, 2023 11:17 am

Doctor offices / insurance





Detailed DNA information is sensitive



Health conditions

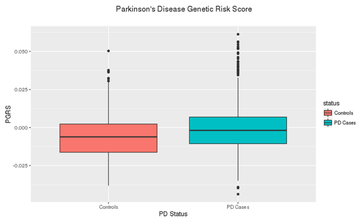
Identity

- Malin, B. and Sweeney, L., 2000. *Determining the identifiability of DNA database entries*. In AMIA Symposium.
- Homer, N., et al. 2008. *Resolving individuals contributing trace amounts of DNA to highly complex mixtures using high-density SNP genotyping microarrays*. PLoS genetics, 4(8).
- Nyholt, D.R., et al. 2009. *On Jim Watson's APOE status: genetic information is hard to hide*. European Journal of Human Genetics, 17(2).
- Wang, R., et al, 2009. *Learning your identity and disease from research papers: information leaks in genome wide association study*. In ACM CCS.
- Gymrek, M., et al. 2013. *Identifying personal genomes by surname inference*. Science, 339(6117).

Lack of PRS privacy guidance

Research publishing, clinical studies, …

	A	B	C	D
1	Cohort	Case control	Disease subtype	Standardized PRS
2	clinical	1	Infiltrating	0.291159147
3	clinical	1	Infiltrating	0.706357435
4	clinical	1	Infiltrating	0.291159147
5	clinical	1	Infiltrating	-0.539237428
6	clinical	1	Infiltrating	0.291159147
7	clinical	1	Infiltrating	0.291159147
8	clinical	1	Infiltrating	1.951952297
9	clinical	1	Infiltrating	0.291159147
10	clinical	1	Infiltrating	1.53675401
11	clinical	1	Infiltrating	1.53675401
12	clinical	1	Infiltrating	-0.539237428
13	clinical	1	Infiltrating	0.706357435



Online platforms

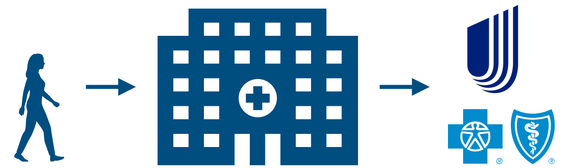
r/genomics • 4 yr. ago
Professional_Lead958

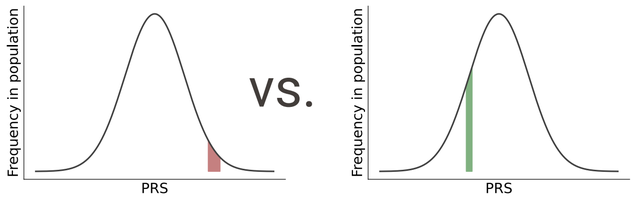
Help! Nebula results

Did Nebula Genomics test and am freaking out — looking for a genetic counselor I can speak with.

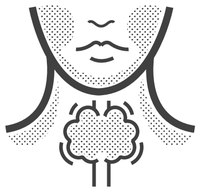
by **Beatrice3** » Sun Dec 31, 2023 11:17 am

Doctor offices / insurance





Detailed DNA information is sensitive



Health conditions

Identity

A G … C … T
T G … G … T

vs. PRS: 1.310749

- Malin, B. and Sweeney, L., 2000. *Determining the identifiability of DNA database entries*. In AMIA Symposium.
- Homer, N., et al. 2008. *Resolving individuals contributing trace amounts of DNA to highly complex mixtures using high-density SNP genotyping microarrays*. PLoS genetics, 4(8).
- Nyholt, D.R., et al. 2009. *On Jim Watson's APOE status: genetic information is hard to hide*. European Journal of Human Genetics, 17(2).
- Wang, R., et al, 2009. *Learning your identity and disease from research papers: information leaks in genome wide association study*. In ACM CCS.
- Gymrek, M., et al. 2013. *Identifying personal genomes by surname inference*. Science, 339(6117).

Graphics from veectezy.com

PRS How-to

$$PRS = \frac{\sum_{i=1}^{N} \beta_i \times X_i}{P \times N}$$

Effect weight β

Number of effect alleles $X \in \{0,1,2\}$

PRS How-to

$$PRS = \frac{\sum_{i=1}^{N} \beta_i \times X_i}{P \times N}$$

Effect weight β

Number of effect alleles $X \in \{0,1,2\}$

```
##POLYGENIC SCORE (PGS) INFORMATION
#pgs_id=PGS000073
#trait_reported=Cervical cancer
#variants_number=10
#weight_type=log(OR)
rsID       chr_name chr_position effect  other.  effect_weight
rs3130196.    6       33063219     C       T     0.3576744442718159

rs3132461.    6       31480668     G       A     0.3074846997479607

rs2523557     6       31331257     G       A     -0.35065687161316933

rs9271775     6       32594328     C       T     0.3293037471426003

rs4713460     6       31347798     G       A     0.3576744442718159

rs2239704.    6       31540141     A       C     -0.20701416938432612

rs9271898.    6       32595972     G       A     0.44468582126144574

rs1882        6       31382911     A       G     0.3576744442718159

rs3132954     6       32311459     G       A     0.3148107398400336

rs73730372.   6       33317843     C       T     -0.5128236264286637
```

PRS How-to

$$PRS = \frac{\sum_{i=1}^{N} \beta_i \times X_i}{P \times N}$$

Effect weight β

Number of effect alleles $X \in \{0,1,2\}$

Commonly high precision β

```
##POLYGENIC SCORE (PGS) INFORMATION
#pgs_id=PGS000073
#trait_reported=Cervical cancer
#variants_number=10
#weight_type=log(OR)
rsID       chr_name chr_position effect  other.  effect_weight
rs3130196.    6      33063219     C       T     0.3576744442718159

rs3132461.    6      31480668     G       A     0.3074846997479607

rs2523557     6      31331257     G       A     -0.35065687161316933

rs9271775     6      32594328     C       T     0.3293037471426003

rs4713460     6      31347798     G       A     0.3576744442718159

rs2239704.    6      31540141     A       C     -0.20701416938432612

rs9271898.    6      32595972     G       A     0.44468582126144574

rs1882        6      31382911     A       G     0.3576744442718159

rs3132954     6      32311459     G       A     0.3148107398400336

rs73730372.   6      33317843     C       T     -0.5128236264286637
```

PRS How-to

$$PRS = \frac{\sum_{i=1}^{N} \beta_i \times X_i}{P \times N}$$

Effect weight β

Number of effect alleles $X \in \{0,1,2\}$

Commonly high precision β

```
##POLYGENIC SCORE (PGS) INFORMATION
#pgs_id=PGS000073
#trait_reported=Cervical cancer
#variants_number=10
#weight_type=log(OR)
rsID       chr_name chr_position effect  other. effect_weight
rs3130196.    6      33063219     C       T     0.3576744442718159
rs3132461.    6      31480668     G       A     0.3074846997479607
rs2523557     6      31331257     G       A     -0.35065687161316933
rs9271775     6      32594328     C       T     0.3293037471426003
rs4713460     6      31347798     G       A     0.3576744442718159
rs2239704.    6      31540141     A       C     -0.20701416938432612
rs9271898.    6      32595972     G       A     0.44468582126144574
rs1882        6      31382911     A       G     0.3576744442718159
rs3132954     6      32311459     G       A     0.3148107398400336
rs73730372.   6      33317843     C       T     -0.5128236264286637
```

PGS002298 (PRS14_esophageal)	PGP000328 » Choi J *et al.* Int J Cancer (2020)	Esophageal cancer	carcinoma of esophagus	14
PGS003387 (best_ESCA_BEEA)	PGP000413 » Namba S *et al.* Cancer Res (2022)	Esophageal adenocarcinoma or Barrett's esophagus	esophageal adenocarcinoma, Barrett esophagus	601,980
PGS003388 (best_ESCA_EA)	PGP000413 » Namba S *et al.* Cancer Res (2022)	Esophageal adenocarcinoma	esophageal adenocarcinoma	356,743
PGS005160 (PRS-ESC)	PGP000711 » Zhu M *et al.* PLoS Med (2025)	Esophageal cancer	carcinoma of esophagus	11

N from single digits to millions

PRS How-to

$$PRS = \frac{\sum_{i=1}^{N} \beta_i \times X_i}{P \times N}$$

Effect weight β

Number of effect alleles $X \in \{0,1,2\}$

Commonly high precision β

```
##POLYGENIC SCORE (PGS) INFORMATION
#pgs_id=PGS000073
#trait_reported=Cervical cancer
#variants_number=10
#weight_type=log(OR)
rsID       chr_name chr_position effect  other. effect_weight
rs3130196.    6      33063219     C       T     0.3576744442718159
rs3132461.    6      31480668     G       A     0.3074846997479607
rs2523557     6      31331257     G       A     -0.35065687161316933
rs9271775     6      32594328     C       T     0.3293037471426003
rs4713460     6      31347798     G       A     0.3576744442718159
rs2239704.    6      31540141     A       C     -0.20701416938432612
rs9271898.    6      32595972     G       A     0.44468582126144574
rs1882        6      31382911     A       G     0.3576744442718159
rs3132954     6      32311459     G       A     0.3148107398400336
rs73730372.   6      33317843     C       T     -0.5128236264286637
```

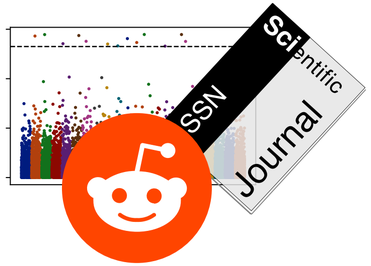
PGS002298 (PRS14_esophageal)	PGP000328 » Choi J *et al.* Int J Cancer (2020)	Esophageal cancer	carcinoma of esophagus	14
PGS003387 (best_ESCA_BEEA)	PGP000413 » Namba S *et al.* Cancer Res (2022)	Esophageal adenocarcinoma or Barrett's esophagus	esophageal adenocarcinoma, Barrett esophagus	601,980
PGS003388 (best_ESCA_EA)	PGP000413 » Namba S *et al.* Cancer Res (2022)	Esophageal adenocarcinoma	esophageal adenocarcinoma	356,743
PGS005160 (PRS-ESC)	PGP000711 » Zhu M *et al.* PLoS Med (2025)	Esophageal cancer	carcinoma of esophagus	11

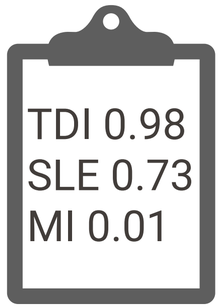
N from single digits to millions

interpretability and cost vs. comprehensive coverage

Attacks vectors

de-identified / anonymous sharing

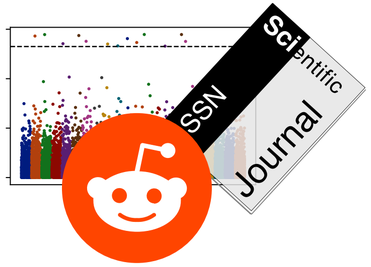




shared with an identity

Attacks vectors

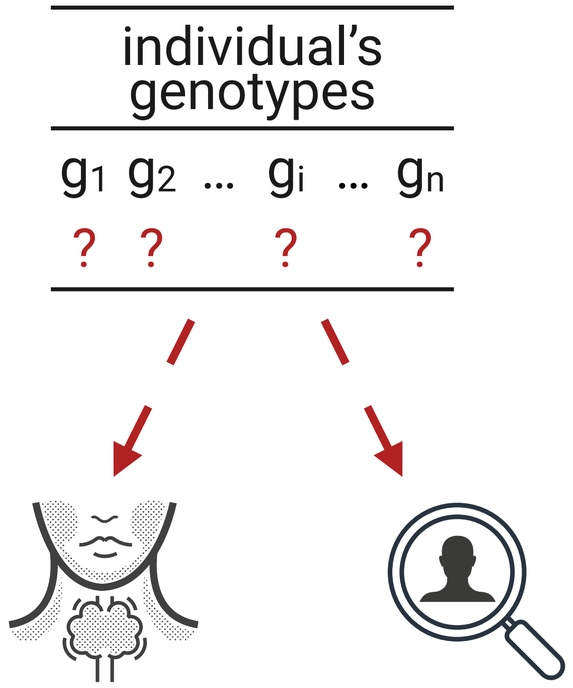
de-identified / anonymous sharing



shared with an identity

?

Can an attacker *recover* genotypes from PRSs and use them to predict other diseases or de-anonymize the individual?



Recovering genotypes from a PRS

Model

rsID	Weight β
rs1	0.374
rs2	0.983
rs3	0.651
rs4	0.553

PRS: 0.34775

rsID	SNP
rs1	?
rs2	?
rs3	?
rs4	?

Recovering genotypes from a PRS

Model

rsID	Weight β
rs1	0.374
rs2	0.983
rs3	0.651
rs4	0.553

PRS: 0.34775

rsID	SNP
rs1	?
rs2	?
rs3	?
rs4	?

$$PRS = \frac{\sum_{i=1}^{N} \beta_i \cdot X_i}{P \cdot N}$$

0.344775 x 4 x 2 = 2.782

− 0.374 x **1** rs1

− 0.983 x **0** rs2

− 0.651 x **2** rs3

− 0.553 x **2** rs4

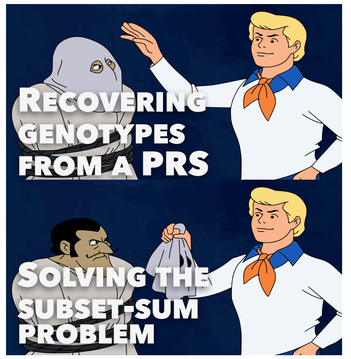
Recovering genotypes from a PRS

Model

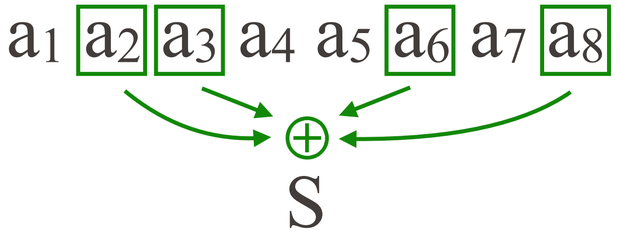
rsID	Weight β
rs1	0.374
rs2	0.983
rs3	0.651
rs4	0.553

PRS: 0.34775

rsID	SNP
rs1	?
rs2	?
rs3	?
rs4	?



The Subset-Sum problem!



$$PRS = \frac{\sum_{i=1}^{N} \beta_i \cdot X_i}{P \cdot N}$$

0.344775 x 4 x 2 = 2.782

− 0.374 x **1** rs1

− 0.983 x **0** rs2

− 0.651 x **2** rs3

− 0.553 x **2** rs4

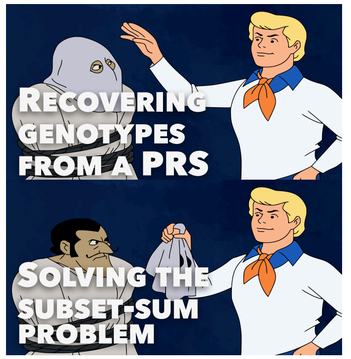
Recovering genotypes from a PRS

Model

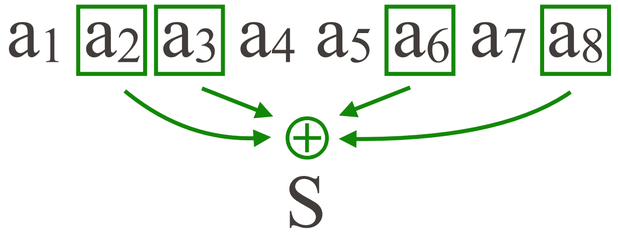
rsID	Weight β
rs1	0.374
rs2	0.983
rs3	0.651
rs4	0.553

PRS: 0.34775

rsID	SNP
rs1	?
rs2	?
rs3	?
rs4	?



The Subset-Sum problem!



$$PRS = \frac{\sum_{i=1}^{N} \beta_i \cdot X_i}{P \cdot N}$$

0.344775 x 4 x 2 = 2.782

− 0.374 x **1** rs1

− 0.983 x **0** rs2

− 0.651 x **2** rs3

− 0.553 x **2** rs4

NP-hard

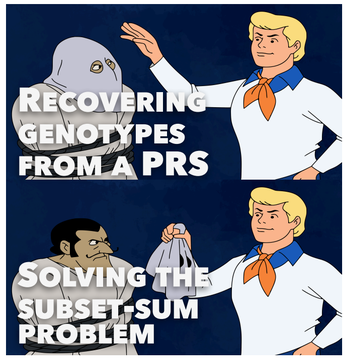
Recovering genotypes from a PRS

Model

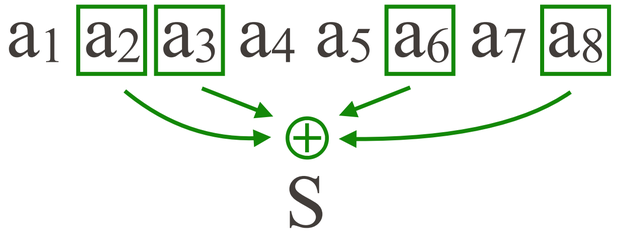
rsID	Weight β
rs1	0.374
rs2	0.983
rs3	0.651
rs4	0.553

PRS: 0.34775

rsID	SNP
rs1	?
rs2	?
rs3	?
rs4	?



The Subset-Sum problem!



$$PRS = \frac{\sum_{i=1}^{N} \beta_i \cdot X_i}{P \cdot N}$$

0.344775 x 4 x 2 = 2.782

− 0.374 x **1** rs1

− 0.983 x **0** rs2

− 0.651 x **2** rs3

− 0.553 x **2** rs4

NP-hard

We can solve it with dynamic programming

Finding solution candidates

PRS
1.03

rsID	Weight
rs1	0.04
rs2	0.11
rs3	0.15
rs4	0.21
rs5	0.38
rs6	0.23
rs7	0.08

Finding solution candidates

PRS
1.03

rsID	Weight
rs1	0.04
rs2	0.11
rs3	0.15
rs4	0.21
rs5	0.38
rs6	0.23
rs7	0.08

Finding solution candidates

PRS
1.03

rsID	Weight
rs1	0.04
rs2	0.11
rs3	0.15
rs4	0.21

rs5	0.38
rs6	0.23
rs7	0.08

Finding solution candidates

PRS
1.03

rsID	Weight
rs1	0.04
rs2	0.11
rs3	0.15
rs4	0.21

0.04	0.08	0.26
0.11	0.22	0.37
0.15	0.30	0.34
0.21	0.42	0.47
...	...	...

0.38
0.72
0.23
0.77
0.61
...

rs5	0.38
rs6	0.23
rs7	0.08

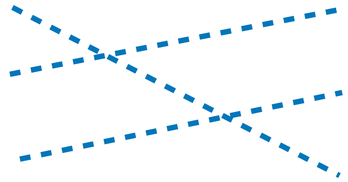
Save the intermediate sum only if it is ≤ target

Finding solution candidates

PRS
1.03

rsID	Weight
rs1	0.04
rs2	0.11
rs3	0.15
rs4	0.21

0.04	0.08	0.26
0.11	0.22	0.37
0.15	0.30	0.34
0.21	0.42	0.47
...	...	...



0.38
0.72
0.23
0.77
0.61
...

rs5	0.38
rs6	0.23
rs7	0.08

Save the intermediate sum only if it is ≤ target

Finding solution candidates

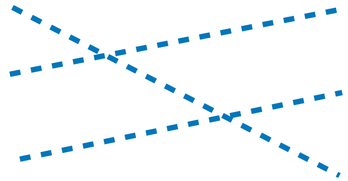
PRS
1.03

rs1	rs2	rs3	rs4
1	2	0	0

rs5	rs6	rs7
1	1	2

rsID	Weight
rs1	0.04
rs2	0.11
rs3	0.15
rs4	0.21

0.04	0.08	0.26
0.11	0.22	0.37
0.15	0.30	0.34
0.21	0.42	0.47
...	...	...



0.38
0.72
0.23
0.77
0.61
...

rs5	0.38
rs6	0.23
rs7	0.08

Save the intermediate sum only if it is ≤ target

Backtrack from the matching subset sums in the end to find all the solutions

Selecting the solution

PRS
1.03

0.04	0.08	0.26
0.11	0.22	0.37
0.15	0.30	0.34
0.21	0.42	0.47
...	...	...

0.38
0.72
0.23
0.77
0.61
...

Selecting the solution

PRS
1.03

0.04	0.08	0.26
0.11	0.22	0.37
0.15	0.30	0.34
0.21	0.42	0.47
...	...	...

0.38
0.72
0.23
0.77
0.61
...

Selecting the solution

PRS
1.03

0.04	0.08	0.26
0.11	0.22	0.37
0.15	0.30	0.34
0.21	0.42	0.47
...	...	...

0.38
0.72
0.23
0.77
0.61
...

Selecting the solution

PRS
1.03

0.04	0.08	0.26
0.11	0.22	0.37
0.15	0.30	0.34
0.21	0.42	0.47
...	...	...

0.38
0.72
0.23
0.77
0.61
...

```
True:
20110220120101120112
Guessed 288:
20110220120101120112 (acc 1.000)
10110220220101120112 (acc 0.900)
20110220120100121112 (acc 0.900)
20110220120102020112 (acc 0.900)
20110220122101100112 (acc 0.900)
20110220120101021112 (acc 0.900)
20110221120001120112 (acc 0.900)
21110220120101120012 (acc 0.900)
20100220120111120112 (acc 0.900)
20110220121101110112 (acc 0.900)
20110220120100220112 (acc 0.900)
20110220120100022112 (acc 0.850)
10110220220102020112 (acc 0.800)
21110220120102020012 (acc 0.800)
21110220120101021012 (acc 0.800)
20100221120011120112 (acc 0.800)
20110220121100111112 (acc 0.800)
10110220220100121112 (acc 0.800)
10110220220100220112 (acc 0.800)
10110221220001120112 (acc 0.800)
20110220122100200112 (acc 0.800)
21110220120100121012 (acc 0.800)
20110220121102010112 (acc 0.800)
```

...

```
11100221222012000012 (acc 0.400)
11100221222011001012 (acc 0.400)
11100221222010101012 (acc 0.400)
11100221221012010012 (acc 0.400)
11100221221010111012 (acc 0.400)
11100221222010002012 (acc 0.350)
11100221221010012012 (acc 0.350)
```

Selecting the solution

PRS
1.03

0.04	0.08	0.26
0.11	0.22	0.37
0.15	0.30	0.34
0.21	0.42	0.47
...	...	...

0.38
0.72
0.23
0.77
0.61
...

```
True:
20110220120101120112
Guessed 288:
20110220120101120112 (acc 1.000)
10110220220101120112 (acc 0.900)
20110220120100121112 (acc 0.900)
20110220120102020112 (acc 0.900)
20110220122101100112 (acc 0.900)
20110220120101021112 (acc 0.900)
20110221120001120112 (acc 0.900)
21110220120101120012 (acc 0.900)
20100220120111120112 (acc 0.900)
20110220121101110112 (acc 0.900)
20110220120100220112 (acc 0.900)
20110220120100022112 (acc 0.850)
10110220220102020112 (acc 0.800)
21110220120102020012 (acc 0.800)
21110220120101021012 (acc 0.800)
20100221120011120112 (acc 0.800)
20110220121100111112 (acc 0.800)
10110220220100121112 (acc 0.800)
10110220220100220112 (acc 0.800)
10110221220001120112 (acc 0.800)
20110220122100200112 (acc 0.800)
21110220120100121012 (acc 0.800)
20110220121102010112 (acc 0.800)
```

...

```
11100221222012000012 (acc 0.400)
11100221222011001012 (acc 0.400)
11100221222010101012 (acc 0.400)
11100221221012010012 (acc 0.400)
11100221221010111012 (acc 0.400)
11100221222010002012 (acc 0.350)
11100221221010012012 (acc 0.350)
```

How do we select *the* solution?

Selecting the solution

SNPs in a human genome are not uniformly random, especially the ones selected for a PRS

```
True:
20110220120101120112
Guessed 288:
20110220120101120112 (acc 1.000)
10110220220101120112 (acc 0.900)
20110220120100121112 (acc 0.900)
20110220120102020112 (acc 0.900)
20110220122101100112 (acc 0.900)
20110220120101021112 (acc 0.900)
20110221120001120112 (acc 0.900)
21110220120101120012 (acc 0.900)
20100220120111120112 (acc 0.900)
20110220121101110112 (acc 0.900)
20110220120100220112 (acc 0.900)
20110220120100022112 (acc 0.850)
10110220220102020112 (acc 0.800)
21110220120102020012 (acc 0.800)
21110220120101021012 (acc 0.800)
20100221120011120112 (acc 0.800)
20110220121100111112 (acc 0.800)
10110220220100121112 (acc 0.800)
10110220220100220112 (acc 0.800)
10110221220001120112 (acc 0.800)
20110220122100200112 (acc 0.800)
21110220120100121012 (acc 0.800)
20110220121102010112 (acc 0.800)
```

...

```
11100221222012000012 (acc 0.400)
11100221222011001012 (acc 0.400)
11100221222010101012 (acc 0.400)
11100221221012010012 (acc 0.400)
11100221221010111012 (acc 0.400)
11100221222010002012 (acc 0.350)
11100221221010012012 (acc 0.350)
```

Selecting the solution

SNPs in a human genome are not uniformly random, especially the ones selected for a PRS

rsID	Allele Freq	
	OTHER	EFFECT
rs1	90%	10%
rs2	53%	47%
rs4	38%	42%

```
True:
20110220120101120112
Guessed 288:
20110220120101120112 (acc 1.000)
10110220220101120112 (acc 0.900)
20110220120100121112 (acc 0.900)
20110220120102020112 (acc 0.900)
20110220122101100112 (acc 0.900)
20110220120101021112 (acc 0.900)
20110221120001120112 (acc 0.900)
21110220120101120012 (acc 0.900)
20100220120111120112 (acc 0.900)
20110220121101110112 (acc 0.900)
20110220120100220112 (acc 0.900)
20110220120100022112 (acc 0.850)
10110220220102020112 (acc 0.800)
21110220120102020012 (acc 0.800)
21110220120101021012 (acc 0.800)
20100221120011120112 (acc 0.800)
20110220121100111112 (acc 0.800)
10110220220100121112 (acc 0.800)
10110220220100220112 (acc 0.800)
10110221220001120112 (acc 0.800)
20110220122100200112 (acc 0.800)
21110220120100121012 (acc 0.800)
20110220121102010112 (acc 0.800)
```

...

```
11100221222012000012 (acc 0.400)
11100221222011001012 (acc 0.400)
11100221222010101012 (acc 0.400)
11100221221012010012 (acc 0.400)
11100221221010111012 (acc 0.400)
11100221222010002012 (acc 0.350)
11100221221010012012 (acc 0.350)
```

Selecting the solution

SNPs in a human genome are not uniformly random, especially the ones selected for a PRS

rsID	Allele Freq	
	OTHER	EFFECT
rs1	90%	10%
rs2	53%	47%
rs4	38%	42%

We can use the likelihood of each solution, given the allele frequency in the population, to rank solution candidates

$$P(g_0, \ldots, g_{n-1}) = \sum_{i=0}^{n-1} \log P(g_i \,|\, AF) = \sum_{i=0}^{n-1} \log P(allele_{i_0}, allele_{i_1} \,|\, AF)$$

```
True:
2011022012010120112
Guessed 288:
2011022012010120112 (acc 1.000)
1011022022010120112 (acc 0.900)
2011022012010012112 (acc 0.900)
2011022012010202012 (acc 0.900)
2011022012210110012 (acc 0.900)
2011022012010102112 (acc 0.900)
2011022112000112012 (acc 0.900)
2111022012010112012 (acc 0.900)
2010022012011112012 (acc 0.900)
2011022012110111012 (acc 0.900)
2011022012010022012 (acc 0.900)
2011022012010002112 (acc 0.850)
1011022022010202012 (acc 0.800)
2111022012010202012 (acc 0.800)
2111022012010102012 (acc 0.800)
2010022112001112012 (acc 0.800)
2011022012110011112 (acc 0.800)
1011022022010012112 (acc 0.800)
1011022022010022012 (acc 0.800)
1011022122000112012 (acc 0.800)
2011022012210020012 (acc 0.800)
2111022012010012012 (acc 0.800)
2011022012110201012 (acc 0.800)
```

...

```
1110022122201200012 (acc 0.400)
1110022122201100102 (acc 0.400)
1110022122201010102 (acc 0.400)
1110022122101201002 (acc 0.400)
1110022122101011102 (acc 0.400)
1110022122201000201 (acc 0.350)
1110022122101001201 (acc 0.350)
```

Selecting the solution

SNPs in a human genome are not uniformly random, especially the ones selected for a PRS

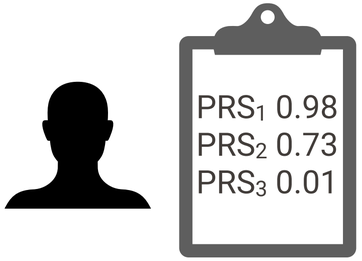
rsID	Allele Freq	
	OTHER	EFFECT
rs1	90%	10%
rs2	53%	47%
rs4	38%	42%

We can use the likelihood of each solution, given the allele frequency in the population, to rank solution candidates

$$P(g_0, \ldots, g_{n-1}) = \sum_{i=0}^{n-1} \log P(g_i \,|\, AF) = \sum_{i=0}^{n-1} \log P(allele_{i_0}, allele_{i_1} \,|\, AF)$$

```
True:
20110220120101120112
Guessed 288:
5.165 20110220120101120112 (acc 1.000)
3.653 10110220220101120112 (acc 0.900)
3.498 20110220120100121112 (acc 0.900)
3.205 20110220120102020112 (acc 0.900)
2.943 20110220122101100112 (acc 0.900)
2.887 20110220120101021112 (acc 0.900)
2.887 20110221120001120112 (acc 0.900)
2.746 21110220120101120012 (acc 0.900)
...   20100220120111120112 (acc 0.900)
      20110220121101110112 (acc 0.900)
      20110220120100220112 (acc 0.900)
      20110220120100022112 (acc 0.850)
      10110220220102020112 (acc 0.800)
      21110220120102020012 (acc 0.800)
      21110220120101021012 (acc 0.800)
      20100221120011120112 (acc 0.800)
      20110220121100111112 (acc 0.800)
      10110220220100121112 (acc 0.800)
      10110220220100220112 (acc 0.800)
      10110221220001120112 (acc 0.800)
      20110220122100200112 (acc 0.800)
      21110220120100121012 (acc 0.800)
      20110220121102010112 (acc 0.800)
...
      11100221222012000012 (acc 0.400)
      11100221222011001012 (acc 0.400)
      11100221222010101012 (acc 0.400)
      11100221221012010012 (acc 0.400)
      11100221221010111012 (acc 0.400)
      11100221222010002012 (acc 0.350)
      11100221221010012012 (acc 0.350)
```

From a single PRS to a panel



Health Predispositions

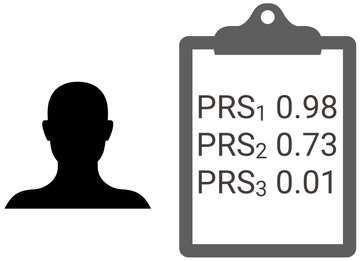
Powered by
23andMe Research

The Health Predispositions ca
meet FDA requirements for G
reports **Powered by 23andM**

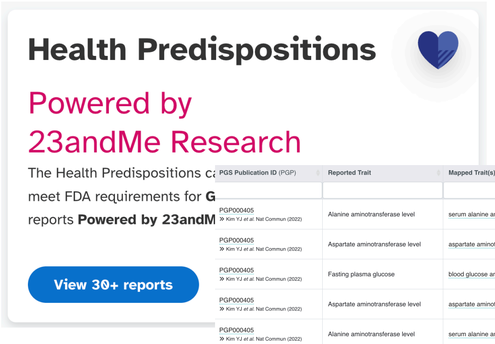
View 30+ reports

PGS Publication ID (PGP)	Reported Trait	Mapped Trait(s) (Ontology)	Number of Variants
PGP000405 » Kim YJ et al. Nat Commun (2022)	Alanine aminotransferase level	serum alanine aminotransferase amount	51
PGP000405 » Kim YJ et al. Nat Commun (2022)	Aspartate aminotransferase level	aspartate aminotransferase measurement	54
PGP000405 » Kim YJ et al. Nat Commun (2022)	Fasting plasma glucose	blood glucose amount	56
PGP000405 » Kim YJ et al. Nat Commun (2022)	Aspartate aminotransferase level	aspartate aminotransferase measurement	56
PGP000405 » Kim YJ et al. Nat Commun (2022)	Alanine aminotransferase level	serum alanine aminotransferase amount	57
PGP000405 » Kim YJ et al. Nat Commun (2022)	Fasting plasma glucose	blood glucose amount	60
PGP000405 » Kim YJ et al. Nat Commun (2022)	LDL cholesterol	low density lipoprotein cholesterol measurement	65
PGP000405 » Kim YJ et al. Nat Commun (2022)	Triglyceride level	triglyceride measurement	65
PGP000405 » Kim YJ et al. Nat Commun (2022)	HbA1c	HbA1c measurement	68
PGP000405 » Kim YJ et al. Nat Commun (2022)	HbA1c	HbA1c measurement	74
PGP000405 » Kim YJ et al. Nat Commun (2022)	Triglyceride level	triglyceride measurement	76

From a single PRS to a panel

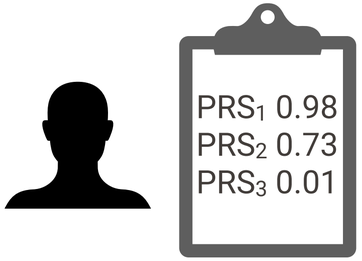


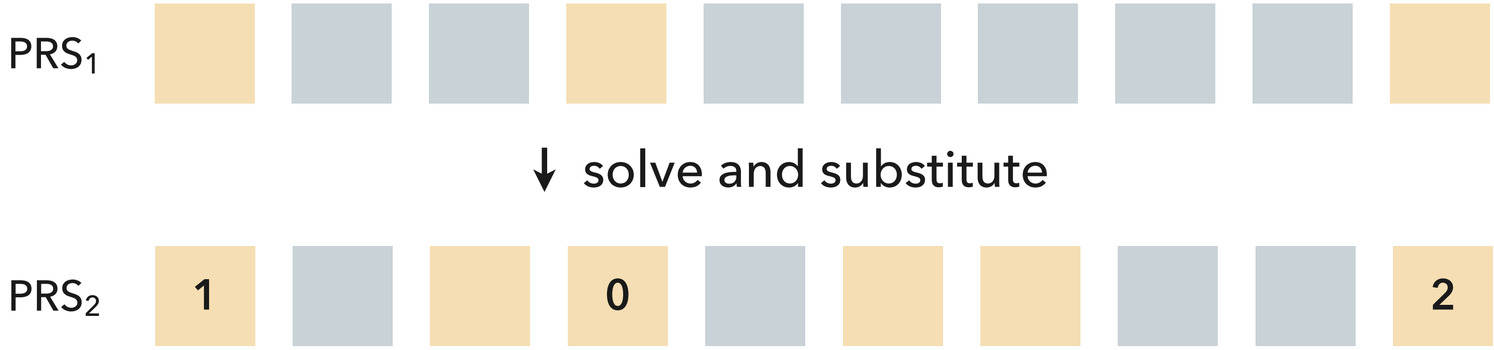
PRS_1

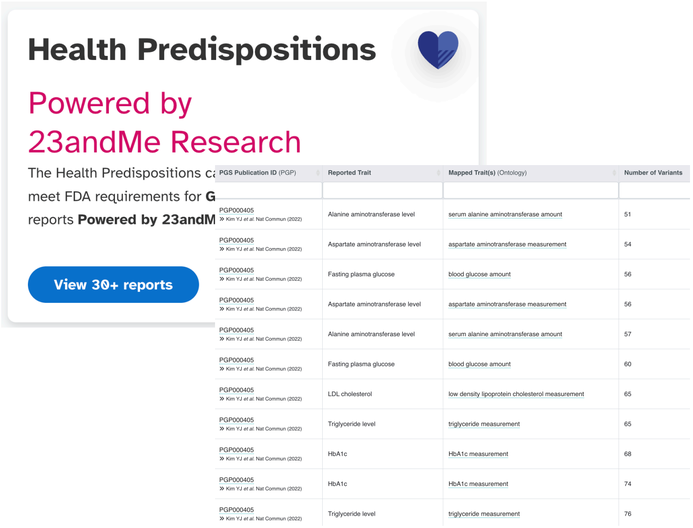


PGS Publication ID (PGP)	Reported Trait	Mapped Trait(s) (Ontology)	Number of Variants
PGP000405 » Kim YJ et al. Nat Commun (2022)	Alanine aminotransferase level	serum alanine aminotransferase amount	51
PGP000405 » Kim YJ et al. Nat Commun (2022)	Aspartate aminotransferase level	aspartate aminotransferase measurement	54
PGP000405 » Kim YJ et al. Nat Commun (2022)	Fasting plasma glucose	blood glucose amount	56
PGP000405 » Kim YJ et al. Nat Commun (2022)	Aspartate aminotransferase level	aspartate aminotransferase measurement	56
PGP000405 » Kim YJ et al. Nat Commun (2022)	Alanine aminotransferase level	serum alanine aminotransferase amount	57
PGP000405 » Kim YJ et al. Nat Commun (2022)	Fasting plasma glucose	blood glucose amount	60
PGP000405 » Kim YJ et al. Nat Commun (2022)	LDL cholesterol	low density lipoprotein cholesterol measurement	65
PGP000405 » Kim YJ et al. Nat Commun (2022)	Triglyceride level	triglyceride measurement	65
PGP000405 » Kim YJ et al. Nat Commun (2022)	HbA1c	HbA1c measurement	68
PGP000405 » Kim YJ et al. Nat Commun (2022)	HbA1c	HbA1c measurement	74
PGP000405 » Kim YJ et al. Nat Commun (2022)	Triglyceride level	triglyceride measurement	76

From a single PRS to a panel







Health Predispositions

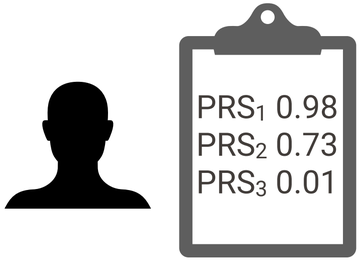
Powered by
23andMe Research

The Health Predispositions ca
meet FDA requirements for G
reports **Powered by 23andM**

View 30+ reports

PGS Publication ID (PGP)	Reported Trait	Mapped Trait(s) (Ontology)	Number of Variants
PGP000405 » Kim YJ et al. Nat Commun (2022)	Alanine aminotransferase level	serum alanine aminotransferase amount	51
PGP000405 » Kim YJ et al. Nat Commun (2022)	Aspartate aminotransferase level	aspartate aminotransferase measurement	54
PGP000405 » Kim YJ et al. Nat Commun (2022)	Fasting plasma glucose	blood glucose amount	56
PGP000405 » Kim YJ et al. Nat Commun (2022)	Aspartate aminotransferase level	aspartate aminotransferase measurement	56
PGP000405 » Kim YJ et al. Nat Commun (2022)	Alanine aminotransferase level	serum alanine aminotransferase amount	57
PGP000405 » Kim YJ et al. Nat Commun (2022)	Fasting plasma glucose	blood glucose amount	60
PGP000405 » Kim YJ et al. Nat Commun (2022)	LDL cholesterol	low density lipoprotein cholesterol measurement	65
PGP000405 » Kim YJ et al. Nat Commun (2022)	Triglyceride level	triglyceride measurement	65
PGP000405 » Kim YJ et al. Nat Commun (2022)	HbA1c	HbA1c measurement	68
PGP000405 » Kim YJ et al. Nat Commun (2022)	HbA1c	HbA1c measurement	74
PGP000405 » Kim YJ et al. Nat Commun (2022)	Triglyceride level	triglyceride measurement	76

From a single PRS to a panel



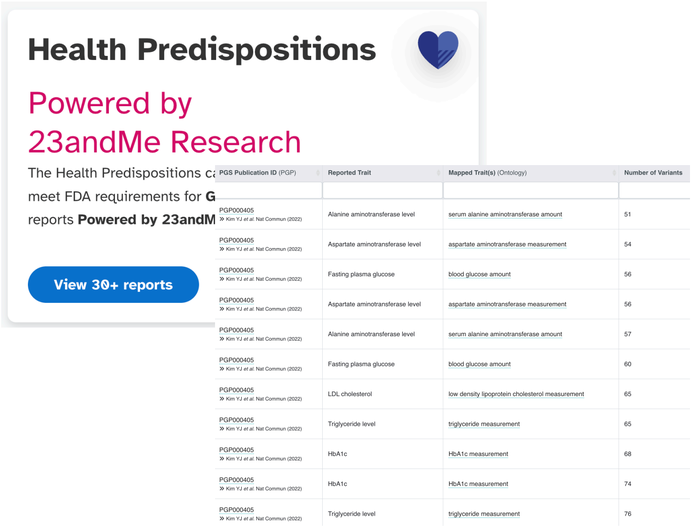
PRS_1

↓ solve and substitute

PRS_2 1 0 2

↓ solve and substitute

PRS_3 1 0 0 2 1 2



Health Predispositions

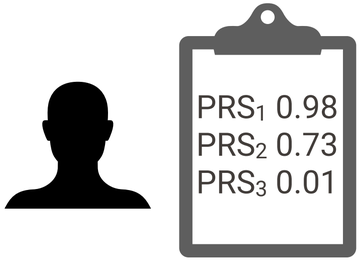
Powered by
23andMe Research

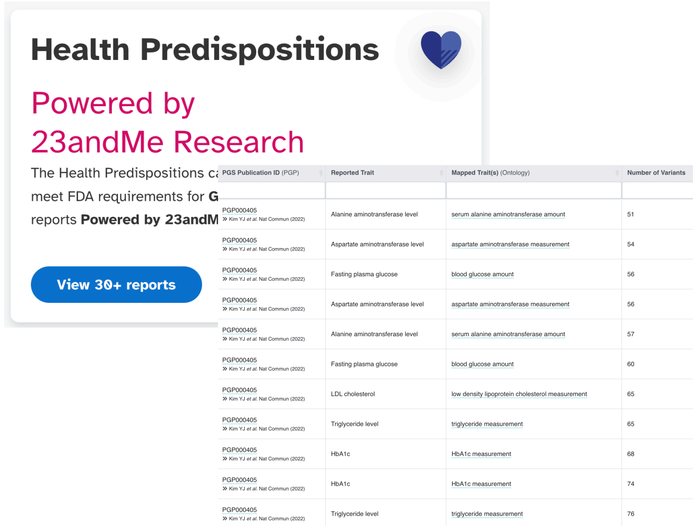
The Health Predispositions ca
meet FDA requirements for **G**
reports **Powered by 23andM**

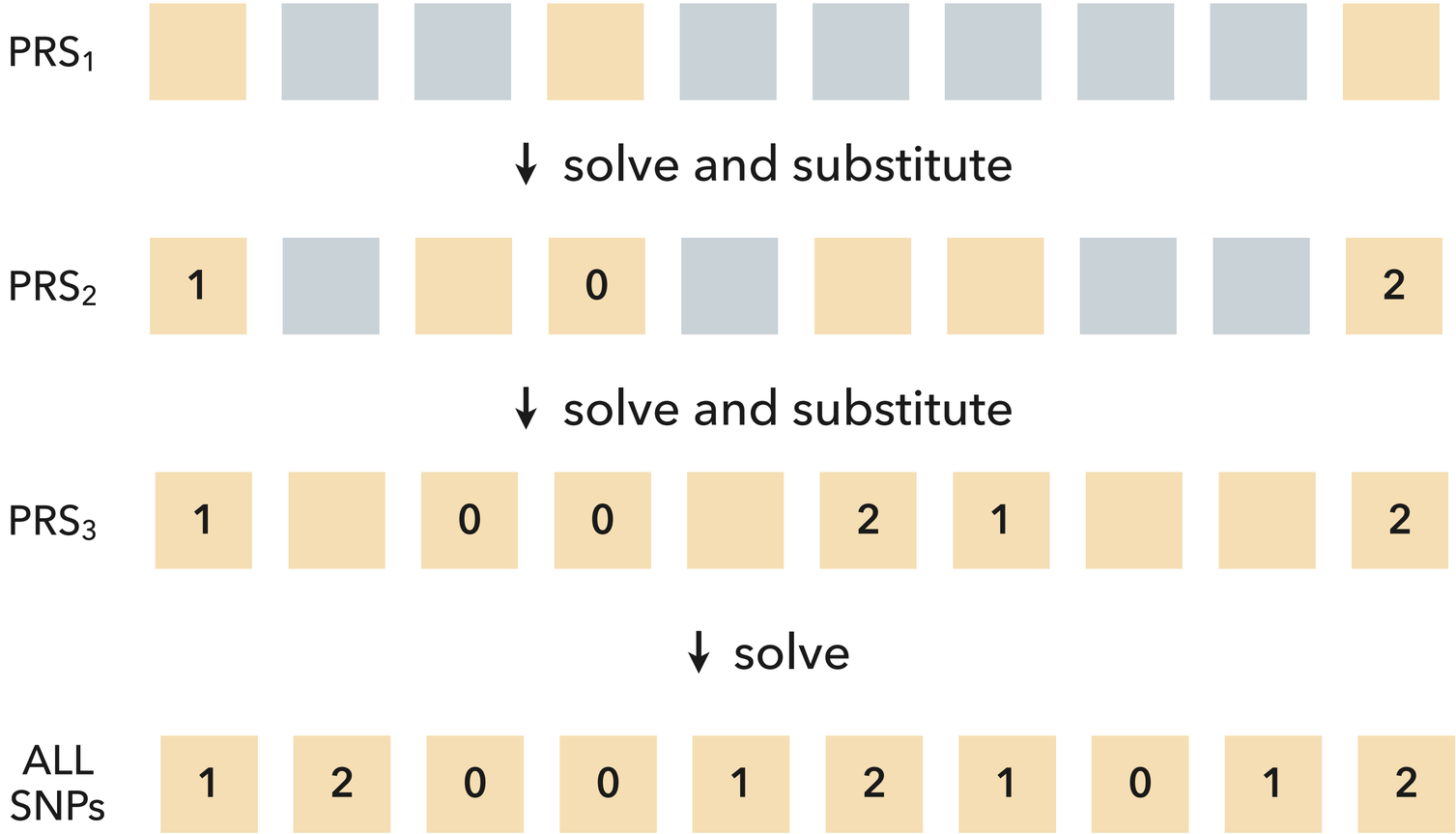
View 30+ reports

PGS Publication ID (PGP)	Reported Trait	Mapped Trait(s) (Ontology)	Number of Variants
PGP000405 » Kim YJ et al. Nat Commun (2022)	Alanine aminotransferase level	serum alanine aminotransferase amount	51
PGP000405 » Kim YJ et al. Nat Commun (2022)	Aspartate aminotransferase level	aspartate aminotransferase measurement	54
PGP000405 » Kim YJ et al. Nat Commun (2022)	Fasting plasma glucose	blood glucose amount	56
PGP000405 » Kim YJ et al. Nat Commun (2022)	Aspartate aminotransferase level	aspartate aminotransferase measurement	56
PGP000405 » Kim YJ et al. Nat Commun (2022)	Alanine aminotransferase level	serum alanine aminotransferase amount	57
PGP000405 » Kim YJ et al. Nat Commun (2022)	Fasting plasma glucose	blood glucose amount	60
PGP000405 » Kim YJ et al. Nat Commun (2022)	LDL cholesterol	low density lipoprotein cholesterol measurement	65
PGP000405 » Kim YJ et al. Nat Commun (2022)	Triglyceride level	triglyceride measurement	65
PGP000405 » Kim YJ et al. Nat Commun (2022)	HbA1c	HbA1c measurement	68
PGP000405 » Kim YJ et al. Nat Commun (2022)	HbA1c	HbA1c measurement	74
PGP000405 » Kim YJ et al. Nat Commun (2022)	Triglyceride level	triglyceride measurement	76

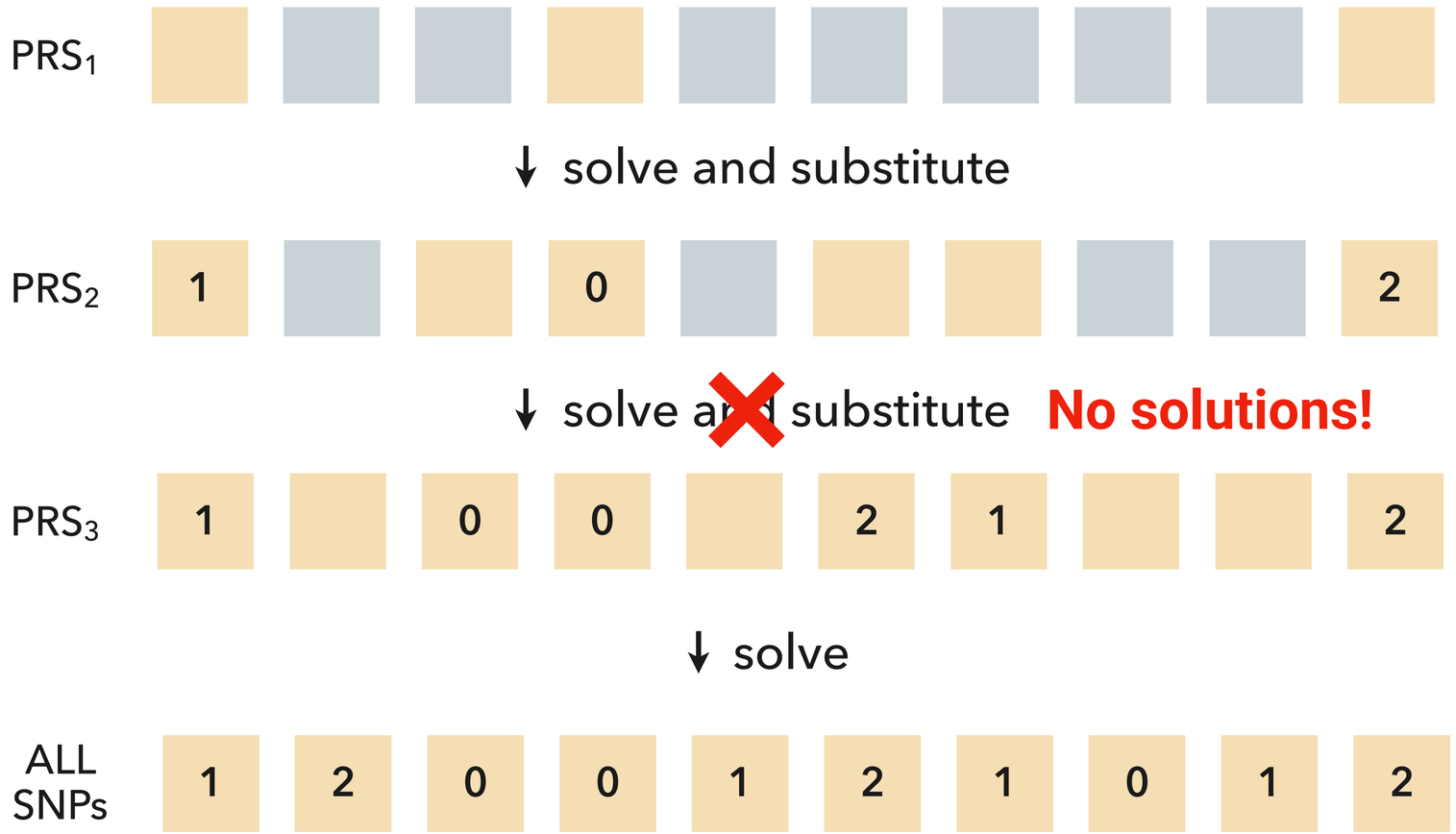
From a single PRS to a panel



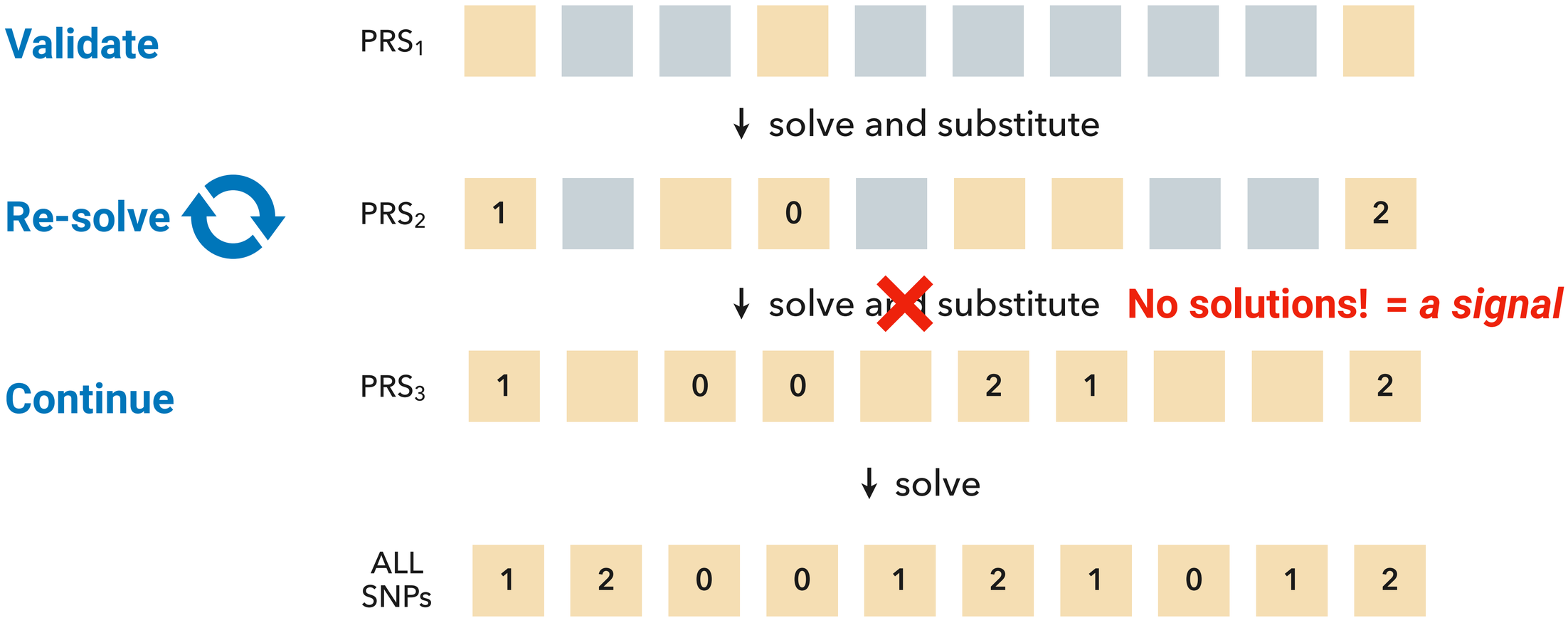




From a single PRS to a panel: Self-repair



From a single PRS to a panel: Self-repair



Solvable PRS

How can an attacker know if a given PRS instance is (easily) solvable?

```
##POLYGENIC SCORE (PGS) INFORMATION
#pgs_id=PGS000073
#trait_reported=Cervical cancer
#variants_number=10
#weight_type=log(OR)
rsID       chr_name chr_position effect  other.  effect_weight
rs3130196.    6      33063219     C       T     0.3576744442718159

rs3132461.    6      31480668     G       A     0.3074846997479607

rs2523557     6      31331257     G       A     -0.35065687161316933

rs9271775     6      32594328     C       T     0.3293037471426003

rs4713460     6      31347798     G       A     0.3576744442718159

rs2239704.    6      31540141     A       C     -0.20701416938432612

rs9271898.    6      32595972     G       A     0.44468582126144574

rs1882        6      31382911     A       G     0.3576744442718159

rs3132954     6      32311459     G       A     0.3148107398400336

rs73730372.   6      33317843     C       T     -0.5128236264286637
```

Solvable PRS

```
##POLYGENIC SCORE (PGS) INFORMATION
#pgs_id=PGS000073
#trait_reported=Cervical cancer
#variants_number=10
#weight_type=log(OR)
rsID       chr_name chr_position effect  other.  effect_weight
rs3130196.    6      33063219     C       T     0.3576744442718159

rs3132461.    6      31480668     G       A     0.3074846997479607

rs2523557     6      31331257     G       A     -0.35065687161316933

rs9271775     6      32594328     C       T     0.3293037471426003

rs4713460     6      31347798     G       A     0.3576744442718159

rs2239704.    6      31540141     A       C     -0.20701416938432612

rs9271898.    6      32595972     G       A     0.44468582126144574

rs1882        6      31382911     A       G     0.3576744442718159

rs3132954     6      32311459     G       A     0.3148107398400336

rs73730372.   6      33317843     C       T     -0.5128236264286637
```

How can an attacker know if a given PRS instance is (easily) solvable?

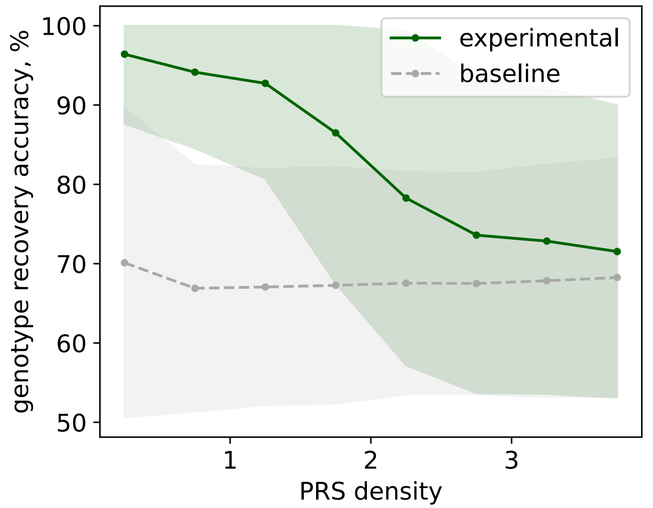
With the concept of subset-sum density

$$d = \frac{N}{\log_3(max_i \text{ decimal}(\beta_i))}$$

Intuitively, a ratio of the total number of possible subsets to the total number of possible sums given the weight precision

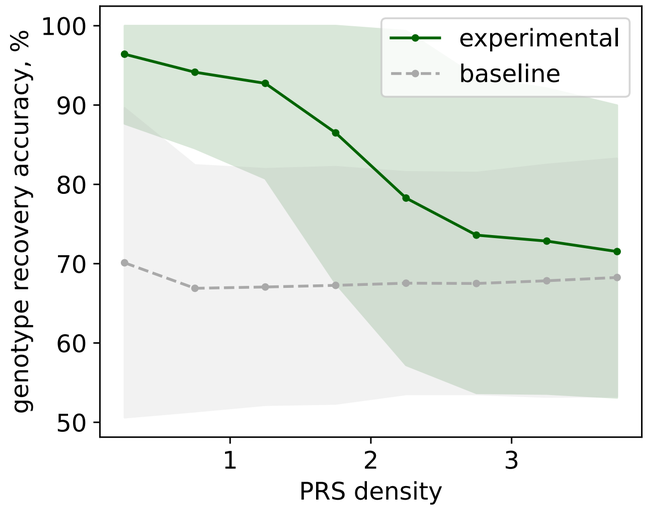
Solvable PRS

Normally, the subset-sum instances with $d < 1$ are considered solvable but we can go up to $d < 2.5$

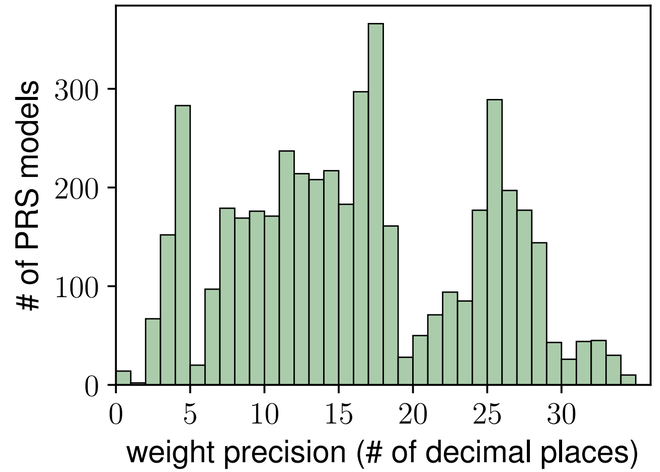


Solvable PRS

Normally, the subset-sum instances with $d < 1$ are considered solvable but we can go up to $d < 2.5$



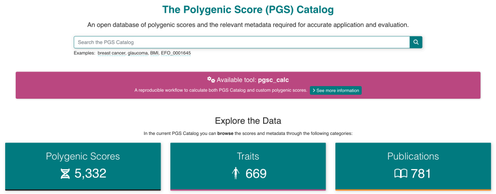
Given the typical precision of effect weights, PRS models with >80 SNPs become too dense for solving reliably



Recovering genotypes: Experimental Setup

- Selecting suitable risk scores from the PGS Catalog

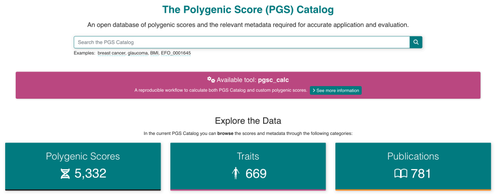
Total PGS with $N < 50$	556
Discarded PGS	
$N = 1$	3
Precision mismatch	3
Mismatching alleles	28
Invalid loci	42
Loci in chr X or Y	51
Density $d > 2.5$	131
Selected PGS	**298**
Total loci	4821
Unique loci	2654



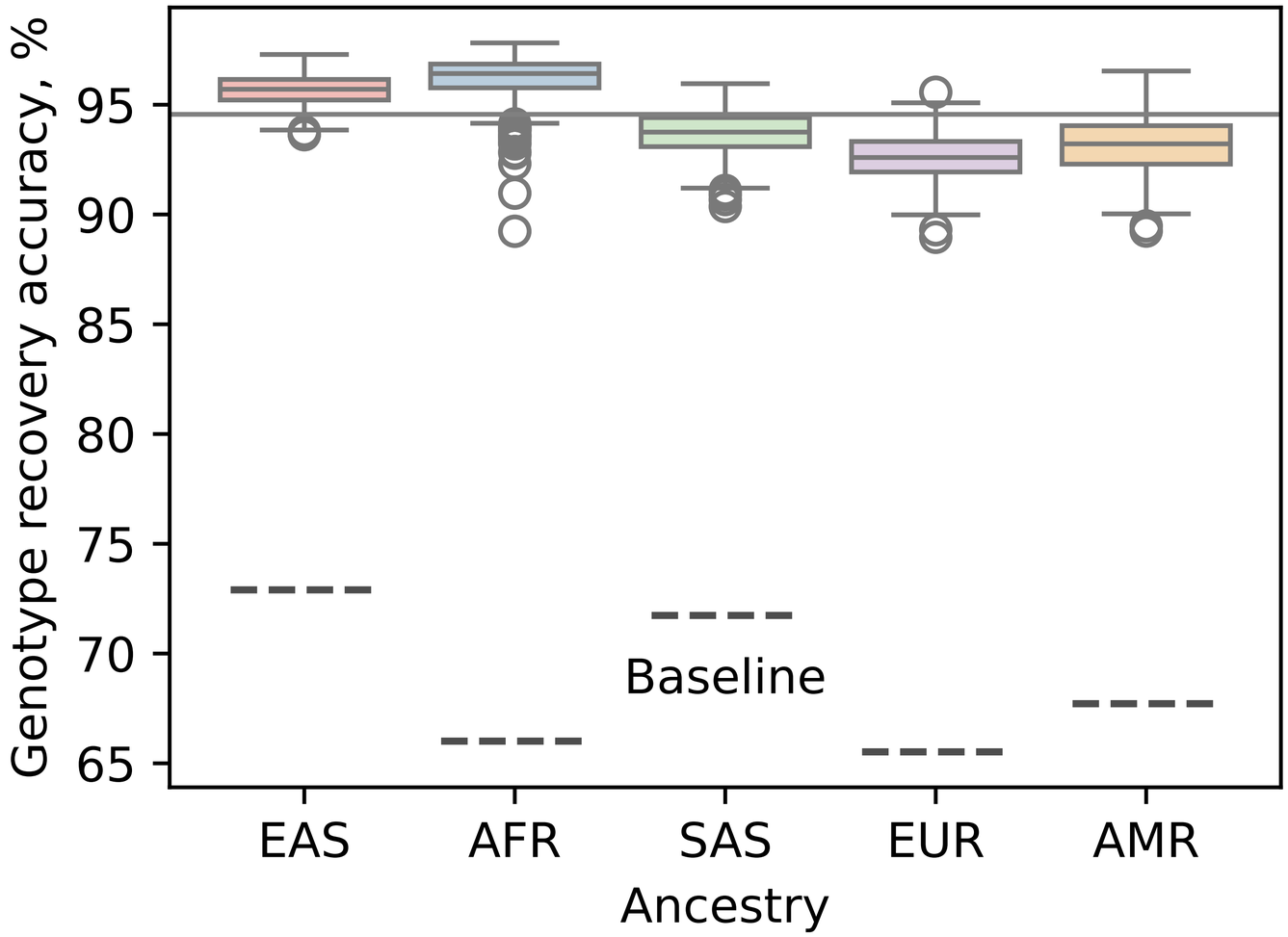
Recovering genotypes: Experimental Setup

- Selecting suitable risk scores from the PGS Catalog
- 2504 primary individuals + 31 relatives from the phase 3 release of the 1000 Genomes dataset (GRCh37)

Total PGS with $N < 50$	556
Discarded PGS	
$N = 1$	3
Precision mismatch	3
Mismatching alleles	28
Invalid loci	42
Loci in chr X or Y	51
Density $d > 2.5$	131
Selected PGS	**298**
Total loci	4821
Unique loci	2654



Recovering genotypes: accuracy



Median number of recovered SNPs: 2600

Median accuracy 94.6%

The baseline is predicting the major genotype for every locus

De-anonymization via a genealogy services

Recovered genotypes

g_1	g_2	g_3	g_4	g_5
0	2	1	0	2
g_6	g_7	g_8	g_9	g_{10}
0	0	1	2	1

De-anonymization via a genealogy services

Recovered genotypes

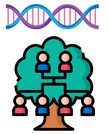
g_1	g_2	g_3	g_4	g_5
0	2	1	0	2
g_6	g_7	g_8	g_9	g_{10}
0	0	1	2	1

De-anonymization via a genealogy services

Recovered genotypes

g_1	g_2	g_3	g_4	g_5
0	2	1	0	2
g_6	g_7	g_8	g_9	g_{10}
0	0	1	2	1

Genetic genealogy database



User	Genotypes
Alice	0|0 1|0 0|1 0|0 1|1 1|0 0|0 1|0 1|1 0|0
Bob	0|0 1|1 1|0 0|0 1|1 0|0 0|0 0|0 1|1 1|0
Carol	1|1 0|1 1|1 0|0 0|0 0|1 0|1 1|1 1|0 1|1

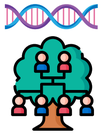
[GED match]

De-anonymization via a genealogy services

Recovered genotypes

g_1	g_2	g_3	g_4	g_5
0	2	1	0	2
g_6	g_7	g_8	g_9	g_{10}
0	0	1	2	1

Submit genotype guesses

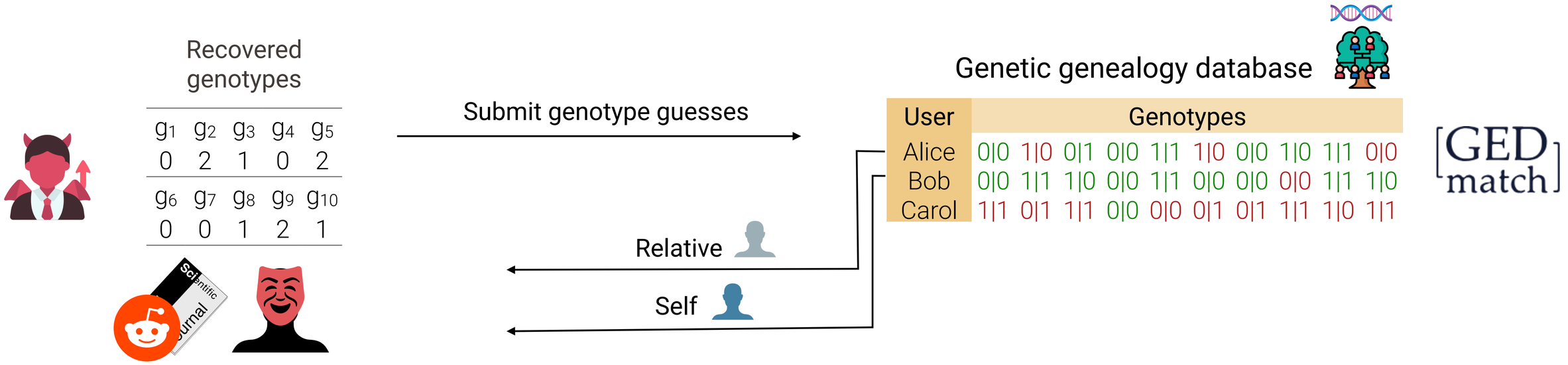


Genetic genealogy database

User	Genotypes
Alice	0|0 1|0 0|1 0|0 1|1 1|0 0|0 1|0 1|1 0|0
Bob	0|0 1|1 1|0 0|0 1|1 0|0 0|0 0|0 1|1 1|0
Carol	1|1 0|1 1|1 0|0 0|0 0|1 0|1 1|1 1|0 1|1

GEDmatch

De-anonymization via a genealogy services



De-anonymization via a genealogy services

Recovered genotypes

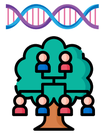
g_1	g_2	g_3	g_4	g_5
0	2	1	0	2
g_6	g_7	g_8	g_9	g_{10}
0	0	1	2	1

Submit genotype guesses

Relative

Self

Genetic genealogy database



User	Genotypes
Alice	0|0 1|0 0|1 0|0 1|1 1|0 0|0 1|0 1|1 0|0
Bob	0|0 1|1 1|0 0|0 1|1 0|0 0|0 0|0 1|1 1|0
Carol	1|1 0|1 1|1 0|0 0|0 0|1 0|1 1|1 1|0 1|1

GEDmatch

KING-robust

$$\hat{\phi}_{ij} = \frac{N_{Aa,Aa} - 2N_{AA,aa}}{N_{Aa}^{(i)} + N_{Aa}^{(j)}}$$

Relationship	ϕ	Inference criteria
Monozygotic twin	$\frac{1}{2}$	$> \frac{1}{2^{3/2}}$
Parent–offspring	$\frac{1}{4}$	$(\frac{1}{2^{5/2}}, \frac{1}{2^{3/2}})$
Full sib	$\frac{1}{4}$	$(\frac{1}{2^{5/2}}, \frac{1}{2^{3/2}})$
2nd Degree	$\frac{1}{8}$	$(\frac{1}{2^{7/2}}, \frac{1}{2^{5/2}})$
3rd Degree	$\frac{1}{16}$	$(\frac{1}{2^{9/2}}, \frac{1}{2^{7/2}})$
Unrelated	0	$< \frac{1}{2^{9/2}}$

Manichaikul, A. et al. *Robust relationship inference in genome-wide association studies*. In Bioinformatics 26 (2010).

De-anonymization via a genealogy services

Recovered genotypes

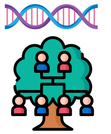
g_1	g_2	g_3	g_4	g_5
0	2	1	0	2
g_6	g_7	g_8	g_9	g_{10}
0	0	1	2	1

Submit genotype guesses

Relative

Self

Genetic genealogy database



User	Genotypes
Alice	0|0 1|0 0|1 0|0 1|1 1|0 0|0 1|0 1|1 0|0
Bob	0|0 1|1 1|0 0|0 1|1 0|0 0|0 0|0 1|1 1|0
Carol	1|1 0|1 1|1 0|0 0|0 0|1 0|1 1|1 1|0 1|1

GEDmatch

KING-robust

$$\hat{\phi}_{ij} = \frac{N_{Aa,Aa} - 2N_{AA,aa}}{N_{Aa}^{(i)} + N_{Aa}^{(j)}}$$

Relationship	ϕ	Inference criteria
Monozygotic twin	$\frac{1}{2}$	$> \frac{1}{2^{3/2}}$
Parent–offspring	$\frac{1}{4}$	$(\frac{1}{2^{5/2}}, \frac{1}{2^{3/2}})$
Full sib	$\frac{1}{4}$	$(\frac{1}{2^{5/2}}, \frac{1}{2^{3/2}})$
2nd Degree	$\frac{1}{8}$	$(\frac{1}{2^{7/2}}, \frac{1}{2^{5/2}})$
3rd Degree	$\frac{1}{16}$	$(\frac{1}{2^{9/2}}, \frac{1}{2^{7/2}})$
Unrelated	0	$< \frac{1}{2^{9/2}}$

Challenges

- The recovered genotypes are *partially* accurate
- Are ~2600 SNPs enough for genetic matching?

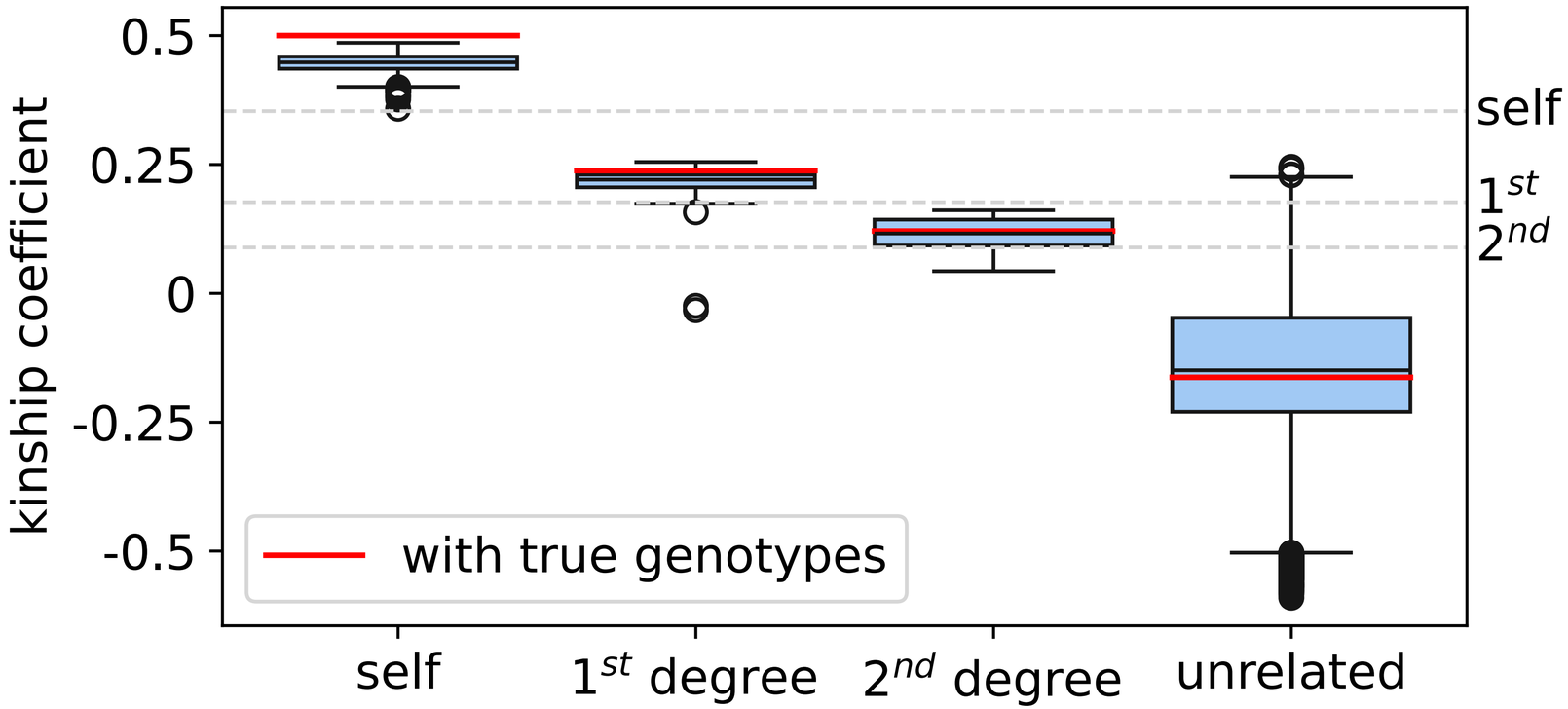
Manichaikul, A. et al. *Robust relationship inference in genome-wide association studies*. In Bioinformatics 26 (2010).

Linking recovered genotypes to 1000 Genomes

2,535 individuals in the database in total

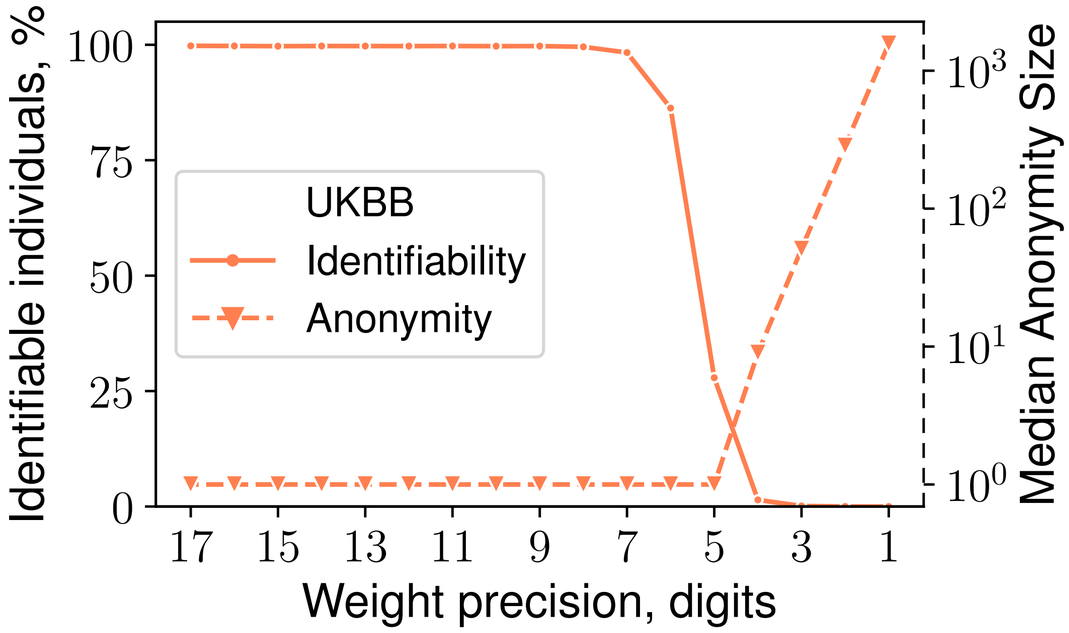
- 50 first-degree relatives
- 18 second-degree

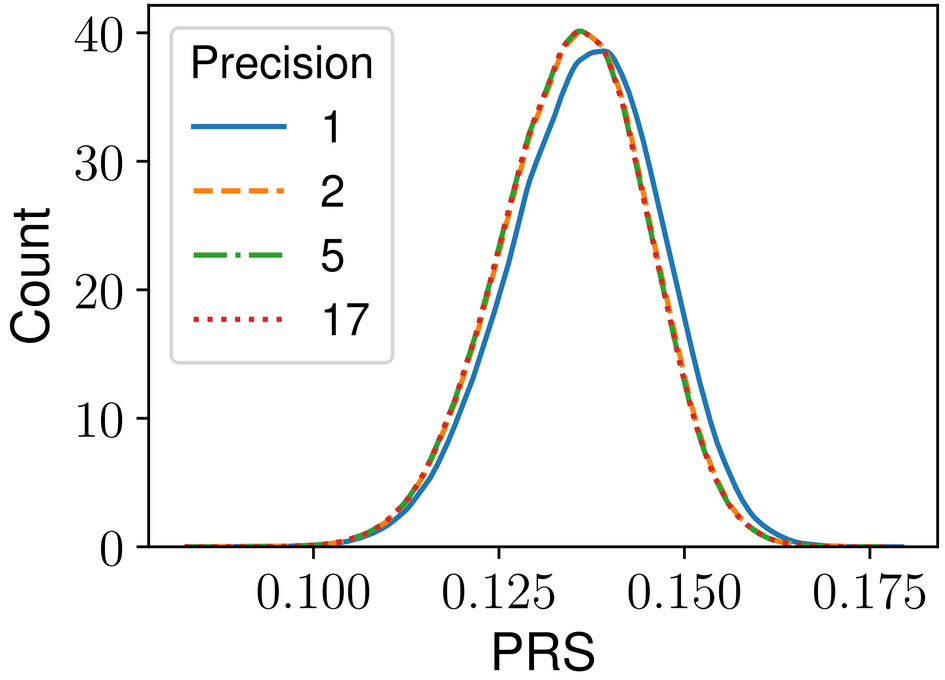
100% precision and recall identifying the individuals themselves



Countermeasures

Decreasing the precision of effect weights reduces identifiability while preserving utility





Conclusion

- Genotypes can be recovered from small-scale PRSs with up to 95% accuracy via dynamic programming and population statistics
- Recovery for large-scale PRSs (>80 SNPs) is challenging due solution density
- 2,600 partially accurate SNPs can re-identify a person in genealogy databases
- Decreasing the precision or adding noise to effect weight improves score privacy

Thanks to the G² lab for all the feedback!

Contact

✉ kirill.nikitin@columbia.edu

𝕏 @ni_kirill