The Polygenic Score Catalog Calculator

A reproducible workflow for PGS calculation

Samuel Lambert[1,2] & Benjamin Wingfield[2]

[1] University of Cambridge [2] EMBL-EBI

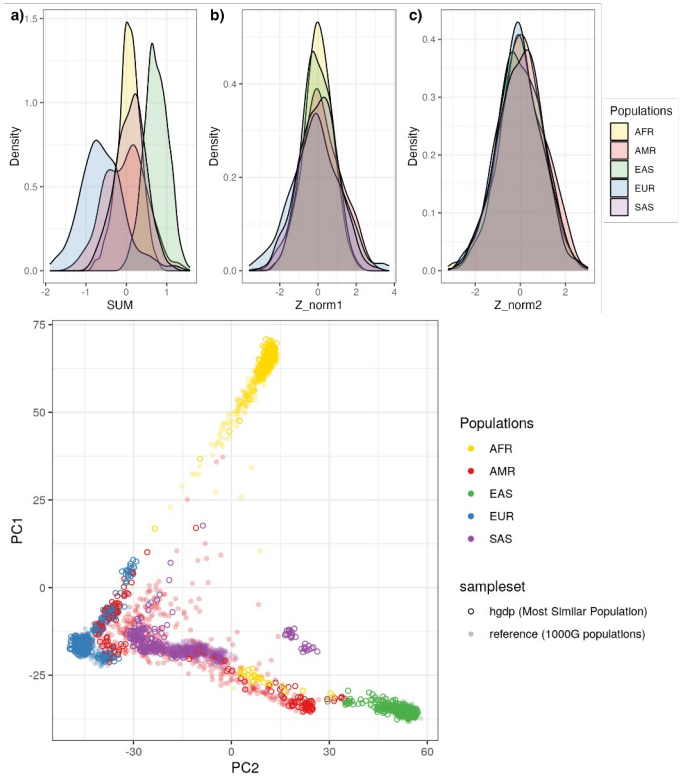
ESHG 2024 Workshop 19.5 14:55 - 15:20

UNIVERSITY OF CAMBRIDGE

EMBL-EBI

What is the PGS Catalog Calculator?

- If you:
  - Have some imputed human genotypes (“target genomes”)
  - Can open a terminal and run some commands
  - Want to calculate some PGS
  - Want to contextualise calculated PGS with genetic ancestry analysis
- Then the PGS Catalog Calculator will be helpful for you



Detailed tutorials available at https://github.com/EBISPOT/eshg-2024-workshop/

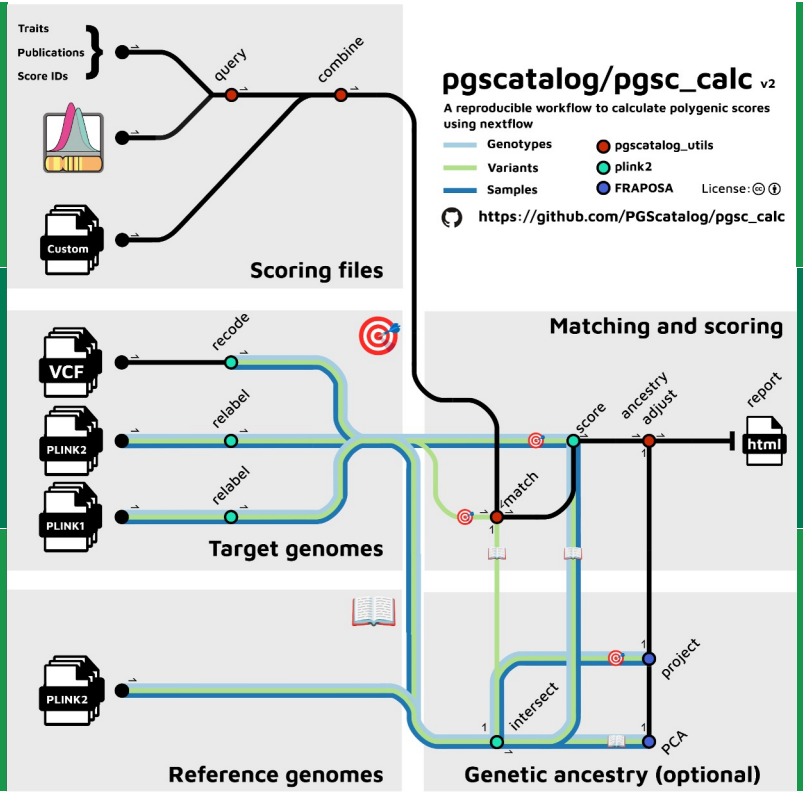
EMBL-EBI

Why is pgsc_calc different from other software?

Bring our code to your data: we're portable!

Deep PGS Catalog integration

Automatically calculate hundreds of PGS in parallel



Robust genetic ancestry analysis and PGS adjustment

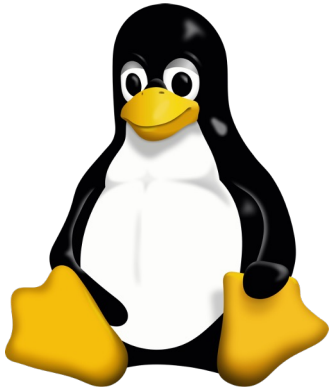
Widely deployed to Trusted Research Environments and biobanks

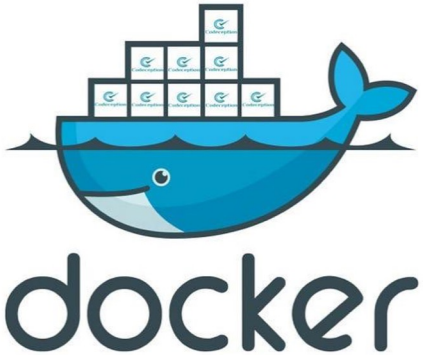
Extensive tests, documentation, and community support

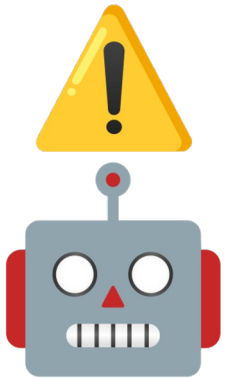
Detailed tutorials available at https://github.com/EBISPOT/eshg-2024-workshop/

EMBL-EBI

Requirements to follow along today







Linux or macOS and an open terminal	Nextflow	Docker Desktop	Admin access (permission to install software)

Detailed tutorials available at https://github.com/EBISPOT/eshg-2024-workshop/

Running the test profile

Fetches code

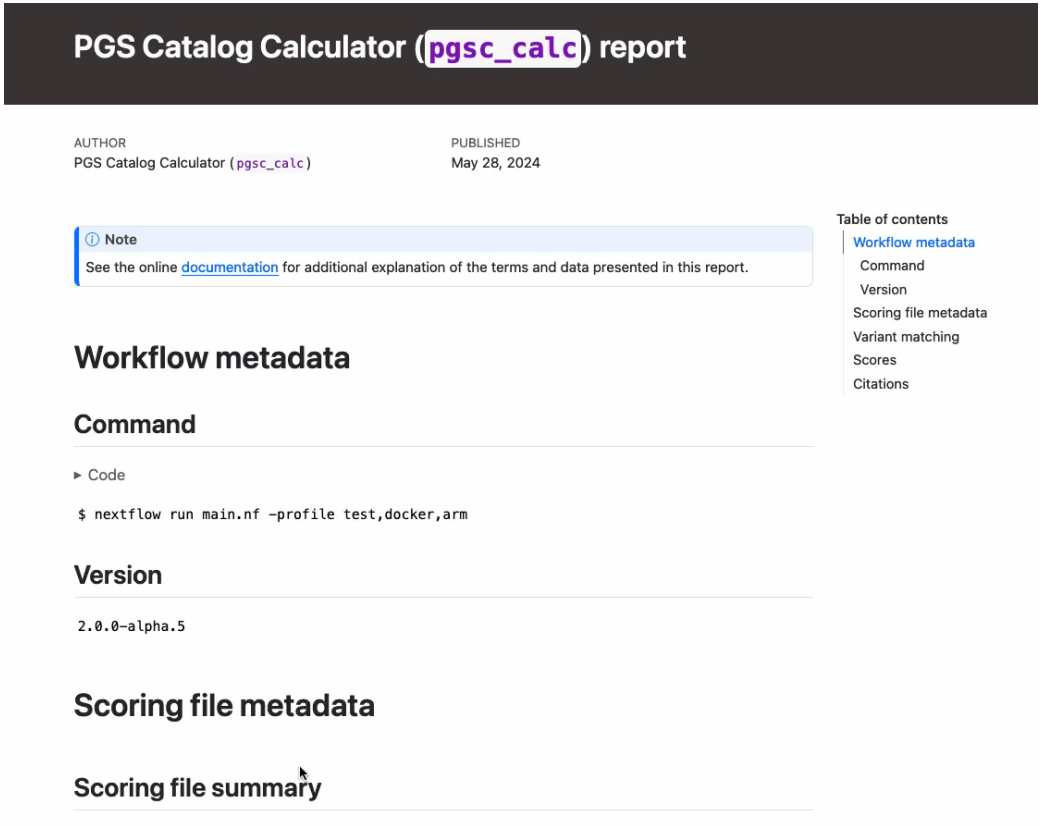
Downloads docker containers

⏰ First download takes longer

Analyses tiny synthetic data

Detailed tutorials available at https://github.com/EBISPOT/eshg-2024-workshop/

Standard output: the report



PGS Catalog Calculator (`pgsc_calc`) report

AUTHOR
PGS Catalog Calculator (`pgsc_calc`)

PUBLISHED
May 28, 2024

Table of contents
- Workflow metadata
  - Command
  - Version
- Scoring file metadata
- Variant matching
- Scores
- Citations

Note
See the online documentation for additional explanation of the terms and data presented in this report.

Workflow metadata

Command

▸ Code

```
$ nextflow run main.nf -profile test,docker,arm
```

Version

```
2.0.0-alpha.5
```

Scoring file metadata

Scoring file summary

Detailed tutorials available at https://github.com/EBISPOT/eshg-2024-workshop/

EMBL-EBI

Standard output: the scores

```
> gzcat aggregated_scores.txt.gz | column -t | head
sampleset  IID      PGS            SUM          DENOM  AVG
cineca     HG00096  PGS001229_22  0.54502      1564   0.0003484782608695652
cineca     HG00097  PGS001229_22  0.674401     1564   0.00043120268542199493
cineca     HG00099  PGS001229_22  0.63727      1564   0.0004074616368286445
cineca     HG00100  PGS001229_22  0.863944     1564   0.0005523938618925831
cineca     HG00101  PGS001229_22  0.280218     1564   0.0001791675191815857
cineca     HG00102  PGS001229_22  0.528136     1564   0.0003376828644501279
cineca     HG00103  PGS001229_22  0.350404     1564   0.00022404347826086957
cineca     HG00105  PGS001229_22  0.51558      1564   0.00032965473145780055
cineca     HG00106  PGS001229_22  0.245779     1564   0.0001571476982097 1867
>
```

Detailed tutorials available at https://github.com/EBISPOT/eshg-2024-workshop/

EMBL-EBI

Standard output: the matching log

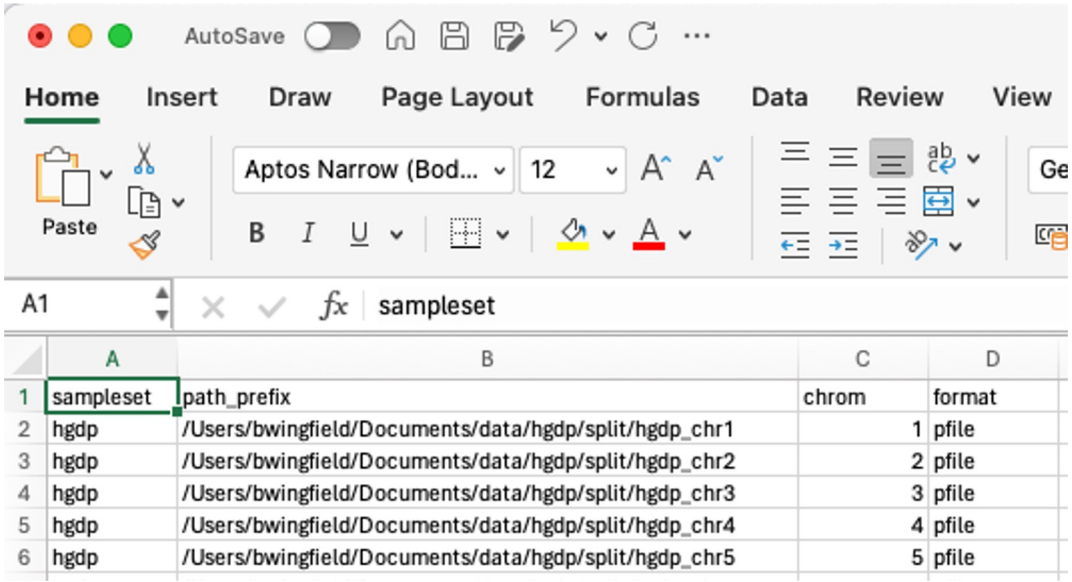
	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W
1	row_nr	accession	chr_name	chr_position	effect_allele	other_allele	effect_weigh	effect_type	ID	REF	ALT	matched_effi	match_type	is_multialleli	ambiguous	match_flippe	best_match	exclude	duplicate_be	duplicate_ID	match_IDs	match_statu	dataset
2	0	PGS001229_	22	17080378	G	A	0.01045457	additive	22:17080378	G	A	G	refalt	FALSE	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	NA	matched	cineca
3	1	PGS001229_	22	17300230	A	G	0.00014115	additive	22:17300230	A	G	A	refalt	FALSE	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	NA	matched	cineca
4	2	PGS001229_	22	17318864	A	C	0.00816627	additive	22:17318864	C	A	A	altref	FALSE	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	NA	matched	cineca
5	3	PGS001229_	22	17327595	T	C	0.00779164	additive	22:17327595	T	C	T	refalt	FALSE	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	NA	matched	cineca
6	4	PGS001229_	22	17409813	A	G	0.00031088	additive	22:17409813	G	A	A	altref	FALSE	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	NA	matched	cineca
7	5	PGS001229_	22	17450952	G	A	-0.0303398	additive	22:17450952	A	G	G	altref	FALSE	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	NA	matched	cineca
8	6	PGS001229_	22	17492533	G	A	0.00388999	additive	22:17492533	G	A	G	refalt	FALSE	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	NA	matched	cineca
9	7	PGS001229_	22	17542810	C	T	0.00803629	additive	22:17542810	C	T	C	refalt	FALSE	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	NA	matched	cineca
10	8	PGS001229_	22	17565013	G	A	0.02135621	additive	22:17565013	G	A	G	refalt	FALSE	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	NA	matched	cineca
11	9	PGS001229_	22	17589209	T	C	0.00302649	additive	22:17589209	C	T	T	altref	FALSE	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	NA	matched	cineca
12	10	PGS001229_	22	17600977	A	G	0.01581277	additive	22:17600977	G	A	A	altref	FALSE	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	NA	matched	cineca
13	11	PGS001229_	22	17625915	A	G	-0.1172964	additive	22:17625915	G	A	A	altref	FALSE	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	NA	matched	cineca
14	12	PGS001229_	22	17630486	A	C	0.01012909	additive	22:17630486	A	C	A	refalt	FALSE	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	NA	matched	cineca
15	13	PGS001229_	22	17633785	C	T	0.0023255	additive	22:17633785	C	T	C	refalt	FALSE	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	NA	matched	cineca
16	14	PGS001229_	22	17643689	A	G	0.00336181	additive	22:17643689	A	G	A	refalt	FALSE	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	NA	matched	cineca
17	15	PGS001229_	22	17669306	C	T	0.0214506	additive	22:17669306	T	C	C	altref	FALSE	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	NA	matched	cineca
18	16	PGS001229_	22	17677699	T	C	-0.0007031	additive	22:17677699	T	C	T	refalt	FALSE	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	NA	matched	cineca
19	17	PGS001229_	22	17680519	C	A	0.00107924	additive	22:17680519	A	C	C	altref	FALSE	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	NA	matched	cineca
21	19	PGS001229_	22	17703119	A	T	0.00077719	additive	22:17703119	T	A	A	altref	FALSE	TRUE	FALSE	TRUE	TRUE	FALSE	FALSE	NA	excluded	cineca
22	19	PGS001229_	22	17703119	A	T	0.00077719	additive	22:17703119	T	A	T	refalt_flip	FALSE	TRUE	TRUE	FALSE	TRUE	FALSE	FALSE	NA	not_best	cineca
24	21	PGS001229_	22	17721595	C	T	0.00948036	additive	22:17721595	T	C	C	altref	FALSE	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	NA	matched	cineca
25	22	PGS001229_	22	17727648	T	C	0.00781169	additive	22:17727648	C	T	T	altref	FALSE	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	NA	matched	cineca
26	23	PGS001229_	22	17738177	G	A	-0.0047198	additive	22:17738177	G	A	G	refalt	FALSE	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	NA	matched	cineca
27	24	PGS001229_	22	17749096	A	G	-0.0052448	additive	22:17749096	G	A	A	altref	FALSE	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	NA	matched	cineca

A list of scoring file variant match candidates. Mostly important for auditing and debugging.

Detailed tutorials available at https://github.com/EBISPOT/eshg-2024-workshop/

EMBL-EBI

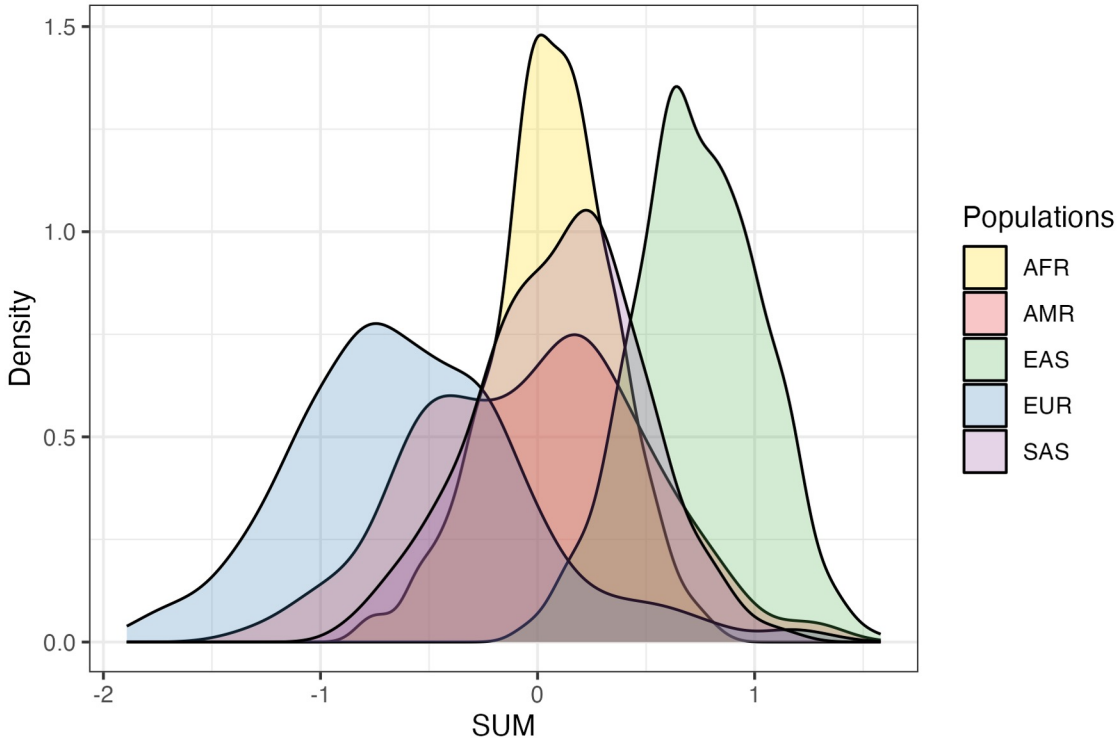
Standard input: the samplesheet



	A	B	C	D
1	sampleset	path_prefix	chrom	format
2	hgdp	/Users/bwingfield/Documents/data/hgdp/split/hgdp_chr1	1	pfile
3	hgdp	/Users/bwingfield/Documents/data/hgdp/split/hgdp_chr2	2	pfile
4	hgdp	/Users/bwingfield/Documents/data/hgdp/split/hgdp_chr3	3	pfile
5	hgdp	/Users/bwingfield/Documents/data/hgdp/split/hgdp_chr4	4	pfile
6	hgdp	/Users/bwingfield/Documents/data/hgdp/split/hgdp_chr5	5	pfile

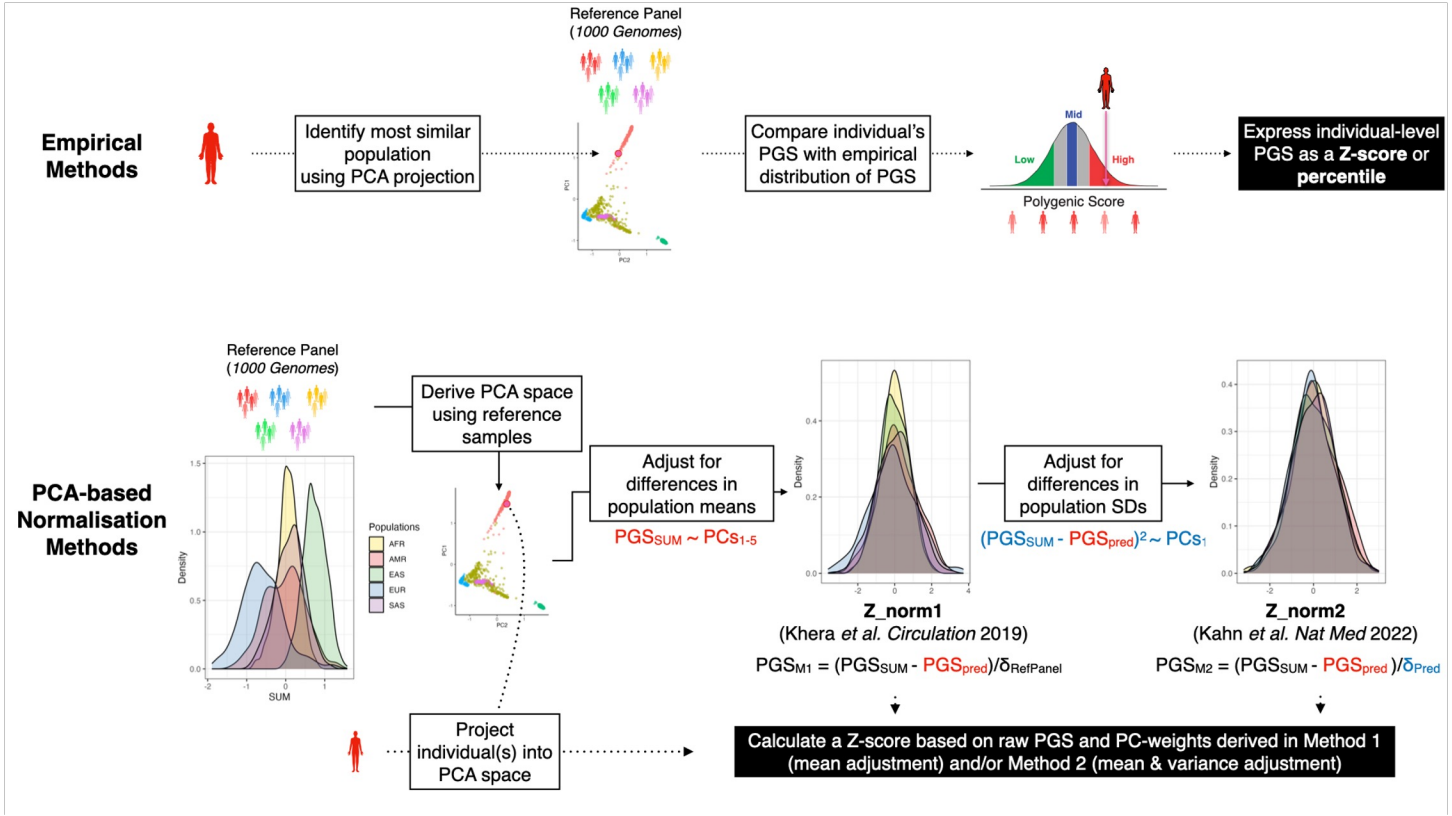
Make a spreadsheet that describes your data (a “samplesheet”)

Why is ancestry adjustment important?



Detailed tutorials available at https://github.com/EBISPOT/eshg-2024-workshop/

Normalisation methods



Detailed tutorials available at https://github.com/EBISPOT/eshg-2024-workshop/

EMBL-EBI

Running ancestry adjustment

Target genomes: HGDP

Reference panel: 1000 Genomes

Calculating PGS001229 (Height)

Adjusting for genetic ancestry

EMBL-EBI

Detailed tutorials available at https://github.com/EBISPOT/eshg-2024-workshop/

Enhanced output: empirically adjusted scores

Empirically adjusted score ⬇️

```
> gzcat hgdp_pgs.txt.gz | cut -f 2,4,5,6,7,8 | column -t | head
IID        SUM                Z_MostSimilarPop        Z_norm1                Z_norm2                percentile_MostSimila
rPop
HGDP00001  3.325961850000001  0.3657432853723631      0.6770295494299402     0.6325425241802616     65.16634050880626
HGDP00003  3.0257247          0.27512800110545277     0.5585400108502155     0.524514602043563      60.86105675146771
HGDP00005  5.6424235000000005  1.0648800541171555      1.3968978974815731     1.3012802716709009     84.93150684931507
HGDP00007  -2.2644397000000005  -1.3215090270727317     -1.0922684992474558    -0.999342637107806     9.197651663405088
HGDP00009  9.7881321          2.316106170794501       2.683036794170566      2.4903331135782087     98.63013698630137
HGDP00011  1.9063992999999997  -0.06269824494095708    0.22598658368718944    0.21563897381449776    48.92367906066536
HGDP00013  2.2672620999999986  0.04621461012495672     0.22560900119684155    0.2235087836755334     53.03326810176125
HGDP00015  3.871218100000001  0.5303083636515148      0.8494048400782984     0.7894233827287974     70.25440313111545
HGDP00017  2.076012770000002  -0.011506802444426487   0.30737436028482135    0.289599901958582      50.88062622309197
>
```

⬆️ Empirically adjusted score

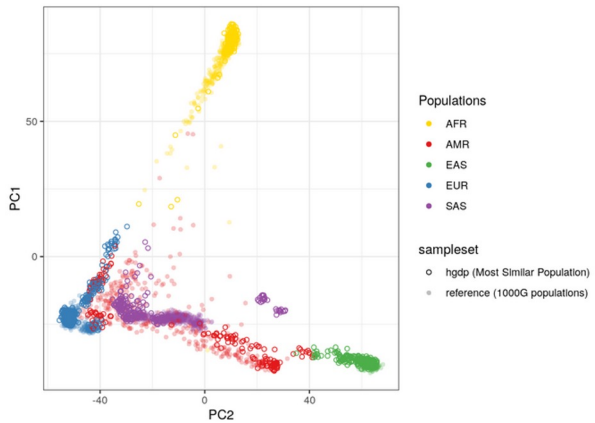
Detailed tutorials available at https://github.com/EBISPOT/eshg-2024-workshop/

EMBL-EBI

Enhanced output: the report

Genetic Ancestry

```
# A tibble: 6 × 22
  sampleset IID          PC1   PC2   PC3    PC4    PC5   PC6    PC7    PC8    PC9
  <chr>     <chr>      <dbl> <dbl> <dbl>  <dbl>  <dbl> <dbl>  <dbl>  <dbl>  <dbl>
1 hgdp      HGDP00001 -18.5 -31.2 -19.7 -2.40  -0.761  5.61  0.973  1.34   2.56
2 hgdp      HGDP00003 -18.8 -29.9 -20.7 -0.866 -0.304  7.31 -0.186  0.350  5.56
3 hgdp      HGDP00005 -18.4 -29.7 -19.1 -4.00  -2.59   5.66  2.67   1.62  -0.907
4 hgdp      HGDP00007 -19.1 -32.1 -23.8 -0.913 -1.43   7.88 -0.120 -1.00   2.32
5 hgdp      HGDP00009 -18.2 -30.8 -21.5 -1.09  -0.721  6.75 -0.241 -0.931  1.09
6 hgdp      HGDP00011 -20.1 -32.1 -21.9 -0.693  0.879  6.62 -0.310  1.42   2.92
# ℹ 11 more variables: PC10 <dbl>, Unrelated <lgl>, RF_P_AFR <dbl>,
#   RF_P_AMR <dbl>, RF_P_EAS <dbl>, RF_P_EUR <dbl>, RF_P_SAS <dbl>,
#   MostSimilarPop <chr>, MostSimilarPop_LowConfidence <lgl>, REFERENCE <lgl>,
#   SuperPop <chr>
```



Population similarity summary

Show 10 entries Search:

	Most similar population	hgdp	reference
1	AFR	110 (11.84%)	681 (26.45%)
2	AMR	165 (17.76%)	351 (13.63%)
3	EAS	209 (22.5%)	507 (19.69%)
4	EUR	251 (27.02%)	525 (20.39%)
5	SAS	194 (20.88%)	511 (19.84%)

Showing 1 to 5 of 5 entries Previous 1 Next

Detailed tutorials available at https://github.com/EBISPOT/eshg-2024-workshop/

EMBL-EBI

Enhanced output: PCA adjusted scores

PCA adjusted score ⬇️

```
> gzcat hgdp_pgs.txt.gz | cut -f 2,4,5,6,7,8 | column -t | head
IID        SUM                 Z_MostSimilarPop         Z_norm1                Z_norm2                percentile_MostSimila
rPop
HGDP00001  3.325961850000001   0.3657432853723631       0.6770295494299402     0.6325425241802616     65.16634050880626
HGDP00003  3.0257247           0.27512800110545277      0.5585400108502155     0.524514602043563      60.86105675146771
HGDP00005  5.6424235000000005  1.0648800541171555       1.3968978974815731     1.3012802716709009     84.93150684931507
HGDP00007  -2.2644397000000005 -1.3215090270727317      -1.0922684992474558    -0.999342637107806     9.197651663405088
HGDP00009  9.7881321           2.316106170794501        2.683036794170566      2.4903331135782087     98.63013698630137
HGDP00011  1.9063992999999997  -0.06269824494095708    0.22598658368718944    0.21563897381449776    48.92367906066536
HGDP00013  2.2672620999999986  0.04621461012495672      0.22560900119684155    0.2235087836755334     53.03326810176125
HGDP00015  3.871218100000001   0.5303083636515148       0.8494048400782984     0.7894233827287974     70.25440313111545
HGDP00017  2.076012770000002   -0.011506802444426487    0.30737436028482135    0.289599901958582      50.88062622309197
>
```

⬆️ PCA adjusted score

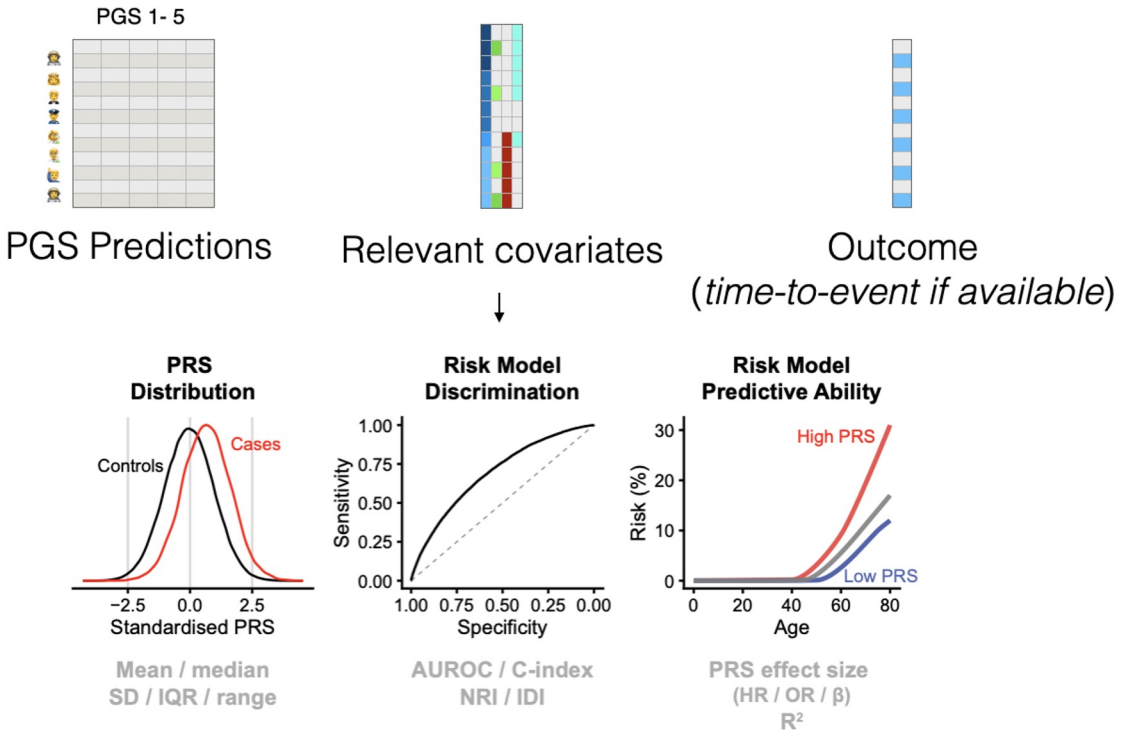
EMBL-EBI

Detailed tutorials available at https://github.com/EBISPOT/eshg-2024-workshop/

So what can you do with the PGS output?

Example use cases:

- Use PGS to model genetic predisposition to different exposures or outcomes
- Use PCs and/or population similarity labels to assess the transferability of PGS



Detailed tutorials available at https://github.com/EBISPOT/eshg-2024-workshop/

EMBL-EBI

Summary

- The PGS Catalog Calculator can portably calculate PGS at biobank scale
- The calculator supports automatic genetic ancestry estimation and adjusts calculated PGS in the context of genetic ancestry

Potential future work

- Maintain scalability as the PGS Catalog data volume grows (CVD currently has 400 million variants)
- WGS target genome support
- Implementing novel methods to normalize PGS in the context of ancestry
- Integrating absolute risk calculations

Detailed tutorials available at https://github.com/EBISPOT/eshg-2024-workshop/

Use EMBL-EBI services? Please give us feedback!

https://www.ebi.ac.uk/about/news/announcements/services-survey-2024/

Survey closes on **Friday** 7th June

EMBL-EBI

I'm interested in	I want to	Where to get help
GWAS Catalog	Learn more	https://www.ebi.ac.uk/gwas/docs https://www.ebi.ac.uk/gwas/docs/related-resources
	Submit data	https://www.ebi.ac.uk/gwas/deposition/
	Get help with submission	gwas-subs@ebi.ac.uk
	Ask a general question	gwas-info@ebi.ac.uk
PGS Catalog	Learn more	https://www.pgscatalog.org/about/
	Submit data or ask a question	pgs-info@ebi.ac.uk
PGS Catalog Calculator	Learn more	https://pgsc-calc.readthedocs.io/en/latest/
	Ask a question or get help	https://github.com/PGScatalog/pgsc_calc

EMBL-EBI

Detailed tutorials available at https://github.com/EBISPOT/eshg-2024-workshop/