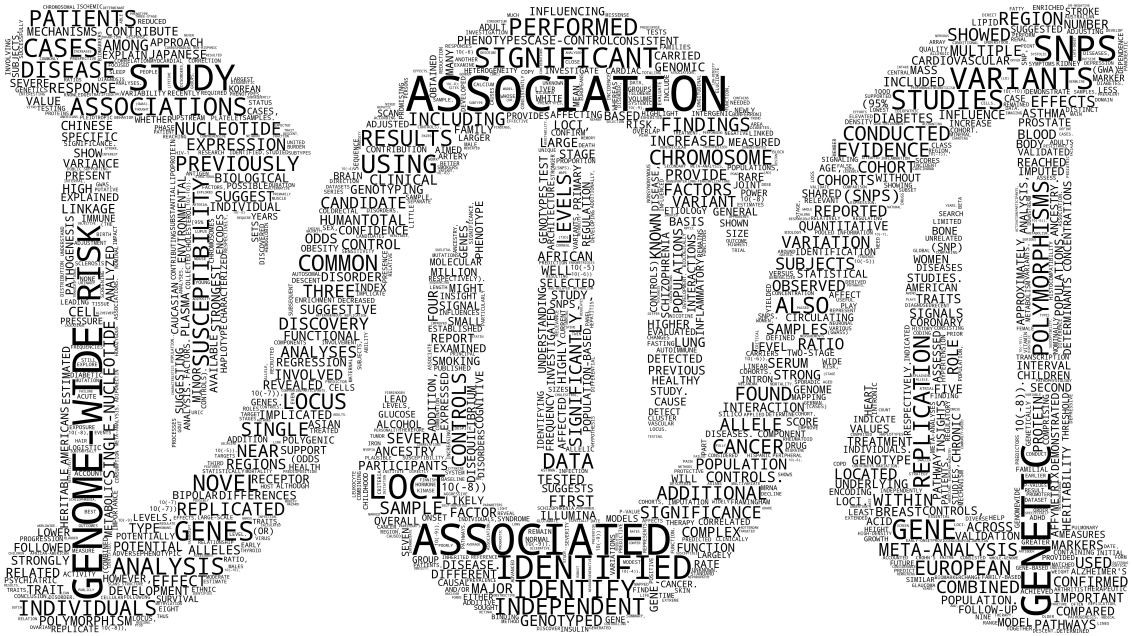
Monitoring GWAS Diversity in Pseudo-Real Time

M.C. Mills, **C. Rahal**, E. Akimova, D. Leasure et al.

Leverhulme Centre for Demographic Science and Nuffield College

Link23 Launch, 29th March, 2023



What is a **G**enome **W**ide **A**ssociation **S**tudy?

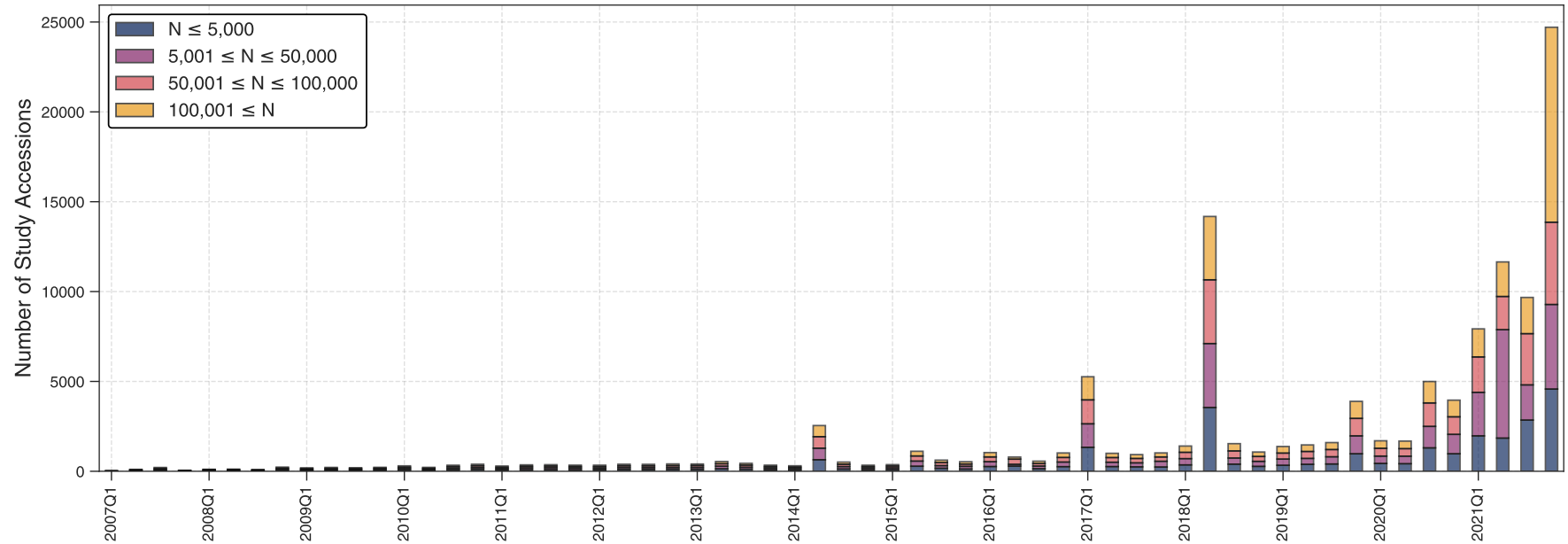
- Observational studies of genome-wide sets of genetic variants.
- Individuals submit DNA to cohort studies, combined into a **GWAS**.
- Scans associations between **S**ingle-**N**ucleotide **P**olymorphisms and traits.
- Requires substantial human and computational capital (and funding).
- Traits can be related to disease (i.e. neoplasty), body measurements (ie anthropometry) or behaviour (i.e. addictions).

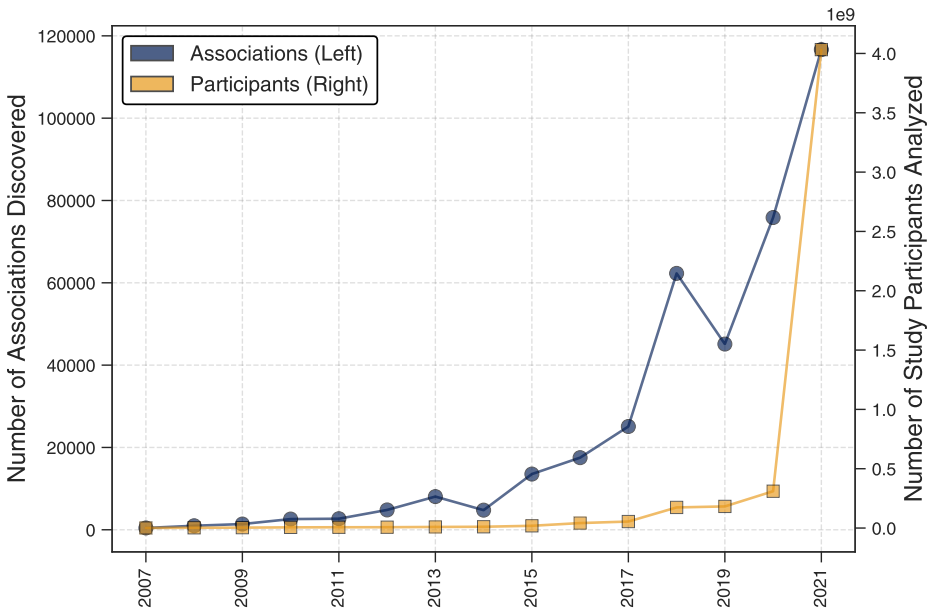
What is a **G**enome **W**ide **A**ssociation **S**tudy?

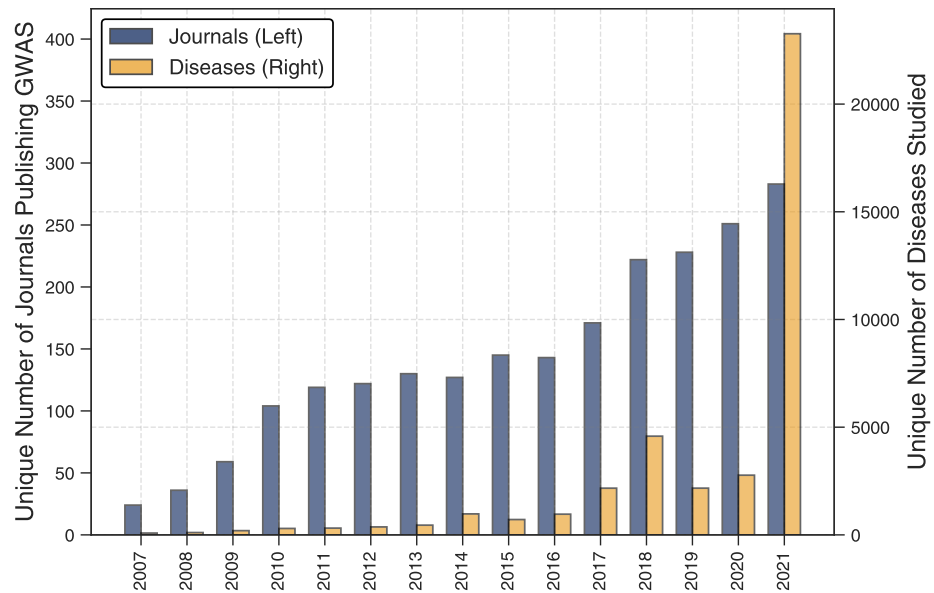
- Observational studies of genome-wide sets of genetic variants.
- Individuals submit DNA to cohort studies, combined into a **GWAS**.
- Scans associations between **S**ingle-**N**ucleotide **P**olymorphisms and traits.
- Requires substantial human and computational capital (and funding).
- Traits can be related to disease (i.e. neoplasty), body measurements (ie anthropometry) or behaviour (i.e. addictions).
- A typical GWAS will have both a *discovery* and *replication* sample.
    - Discovery sample summary statistics generate PRS for clinical purposes.

What is a **G**enome **W**ide **A**ssociation **S**tudy?

- Observational studies of genome-wide sets of genetic variants.
- Individuals submit DNA to cohort studies, combined into a **GWAS**.
- Scans associations between **S**ingle-**N**ucleotide **P**olymorphisms and traits.
- Requires substantial human and computational capital (and funding).
- Traits can be related to disease (i.e. neoplasty), body measurements (ie anthropometry) or behaviour (i.e. addictions).
- A typical GWAS will have both a *discovery* and *replication* sample.
  - Discovery sample summary statistics generate PRS for clinical purposes.
- Specific examples include:
  - **Colorectal cancer**: UK Biobank, 26.3%- 30.1% variance explained.
  - **Schizophrenia**: up to 22% of variance explained.


N ≤ 5,000
5,001 ≤ N ≤ 50,000
50,001 ≤ N ≤ 100,000
100,001 ≤ N
Number of Study Accessions
25000
20000
15000
10000
5000
0
2007Q1
2008Q1
2009Q1
2010Q1
2011Q1
2012Q1
2013Q1
2014Q1
2015Q1
2016Q1
2017Q1
2018Q1
2019Q1
2020Q1
2021Q1


Associations (Left)
Participants (Right)
1e9
Number of Associations Discovered
Number of Study Participants Analyzed
120000
100000
80000
60000
40000
20000
0
4.0
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
2007
2009
2011
2013
2015
2017
2019
2021


Journals (Left)
Diseases (Right)
Unique Number of Journals Publishing GWAS
Unique Number of Diseases Studied
400
350
300
250
200
150
100
50
0
20000
15000
10000
5000
0
2007
2008
2009
2010
2011
2012
2013
2014
2015
2016
2017
2018
2019
2020
2021

Some Basic Facts About GWAS (continually evolving)

- There are currently about **6,306** GWAS papers published (first 2005-03-10).

Some Basic Facts About GWAS (continually evolving)

- There are currently about **6,306** GWAS papers published (first 2005-03-10).
- However: only **10** papers were published in 2005 and 2006 combined.

Some Basic Facts About GWAS (continually evolving)

- There are currently about **6,306** GWAS papers published (first 2005-03-10).
- However: only **10** papers were published in 2005 and 2006 combined.
- Represents around **70,504** Accessions across **55,927** unique Diseases/Traits.

Some Basic Facts About GWAS (continually evolving)

- There are currently about **6,306** GWAS papers published (first 2005-03-10).
- However: only **10** papers were published in 2005 and 2006 combined.
- Represents around **70,504** Accessions across **55,927** unique Diseases/Traits.
- The total number of associations found is currently: **493,105**.

Some Basic Facts About GWAS (continually evolving)

- There are currently about **6,306** GWAS papers published (first 2005-03-10).
- However: only **10** papers were published in 2005 and 2006 combined.
- Represents around **70,504** Accessions across **55,927** unique Diseases/Traits.
- The total number of associations found is currently: **493,105**.
- The average number of associations found per paper is **7**.

Some Basic Facts About GWAS (continually evolving)

- There are currently about **6,306** GWAS papers published (first 2005-03-10).
- However: only **10** papers were published in 2005 and 2006 combined.
- Represents around **70,504** Accessions across **55,927** unique Diseases/Traits.
- The total number of associations found is currently: **493,105**.
- The average number of associations found per paper is **7**.
- The number of associations reaching the 5e-8 threshold is **381,913**.

Some Basic Facts About GWAS (continually evolving)

- There are currently about **6,306** GWAS papers published (first 2005-03-10).
- However: only **10** papers were published in 2005 and 2006 combined.
- Represents around **70,504** Accessions across **55,927** unique Diseases/Traits.
- The total number of associations found is currently: **493,105**.
- The average number of associations found per paper is **7**.
- The number of associations reaching the 5e-8 threshold is **381,913**.
- Journal featuring most GWAS since 2005-03-10: **Nature Communications**.

Some Basic Facts About GWAS (continually evolving)

- There are currently about **6,306** GWAS papers published (first 2005-03-10).
- However: only **10** papers were published in 2005 and 2006 combined.
- Represents around **70,504** Accessions across **55,927** unique Diseases/Traits.
- The total number of associations found is currently: **493,105**.
- The average number of associations found per paper is **7**.
- The number of associations reaching the 5e-8 threshold is **381,913**.
- Journal featuring most GWAS since 2005-03-10: **Nature Communications**.
- This year so far, it's **Nature Medicine** (49 accessions published).

Some Basic Facts About GWAS (continually evolving)

- There are currently about **6,306** GWAS papers published (first 2005-03-10).
- However: only **10** papers were published in 2005 and 2006 combined.
- Represents around **70,504** Accessions across **55,927** unique Diseases/Traits.
- The total number of associations found is currently: **493,105**.
- The average number of associations found per paper is **7**.
- The number of associations reaching the 5e-8 threshold is **381,913**.
- Journal featuring most GWAS since 2005-03-10: **Nature Communications**.
- This year so far, it's **Nature Medicine** (49 accessions published).
- The study with the largest number of authors has **622** authors.

Some Basic Facts About GWAS (continually evolving)

- There are currently about **6,306** GWAS papers published (first 2005-03-10).
- However: only **10** papers were published in 2005 and 2006 combined.
- Represents around **70,504** Accessions across **55,927** unique Diseases/Traits.
- The total number of associations found is currently: **493,105**.
- The average number of associations found per paper is **7**.
- The number of associations reaching the 5e-8 threshold is **381,913**.
- Journal featuring most GWAS since 2005-03-10: **Nature Communications**.
- This year so far, it's **Nature Medicine** (49 accessions published).
- The study with the largest number of authors has **622** authors.
- Largest Accession: **4.1m** people (human height, PMID: 36224396).

The Problem

- PRS provide predictive power above combined clinical risk factors for many traits in **European** populations.

The Problem

- PRS provide predictive power above combined clinical risk factors for many traits in **European** populations.
- However, genetic results discovered in one ancestry are not 'portable':
  - PRS have attenuated cross-population predictive performance.
  - Reductions in portability even in subgroups of European ancestries.
  - Shown in multiple empirical studies, due to genetic architecture, allele frequency, LD patterns, phenotype definition, GxE, etc.

The Problem

- PRS provide predictive power above combined clinical risk factors for many traits in **European** populations.
- However, genetic results discovered in one ancestry are not 'portable':
  - PRS have attenuated cross-population predictive performance.
  - Reductions in portability even in subgroups of European ancestries.
  - Shown in multiple empirical studies, due to genetic architecture, allele frequency, LD patterns, phenotype definition, GxE, etc.
- This is a big problem, because [almost] all large datasets are European.

The Problem

- PRS provide predictive power above combined clinical risk factors for many traits in **European** populations.
- However, genetic results discovered in one ancestry are not 'portable':
  - PRS have attenuated cross-population predictive performance.
  - Reductions in portability even in subgroups of European ancestries.
  - Shown in multiple empirical studies, due to genetic architecture, allele frequency, LD patterns, phenotype definition, GxE, etc.
- This is a big problem, because [almost] all large datasets are European.
- If 'discovering' in European ancestry, benefits don't 'replicate' out-of-sample.

The Problem

- PRS provide predictive power above combined clinical risk factors for many traits in **European** populations.
- However, genetic results discovered in one ancestry are not 'portable':
  - PRS have attenuated cross-population predictive performance.
  - Reductions in portability even in subgroups of European ancestries.
  - Shown in multiple empirical studies, due to genetic architecture, allele frequency, LD patterns, phenotype definition, GxE, etc.
- This is a big problem, because [almost] all large datasets are European.
- If 'discovering' in European ancestry, benefits don't 'replicate' out-of-sample.
- This reduces clinical value in external populations, exacerbating inequalities.

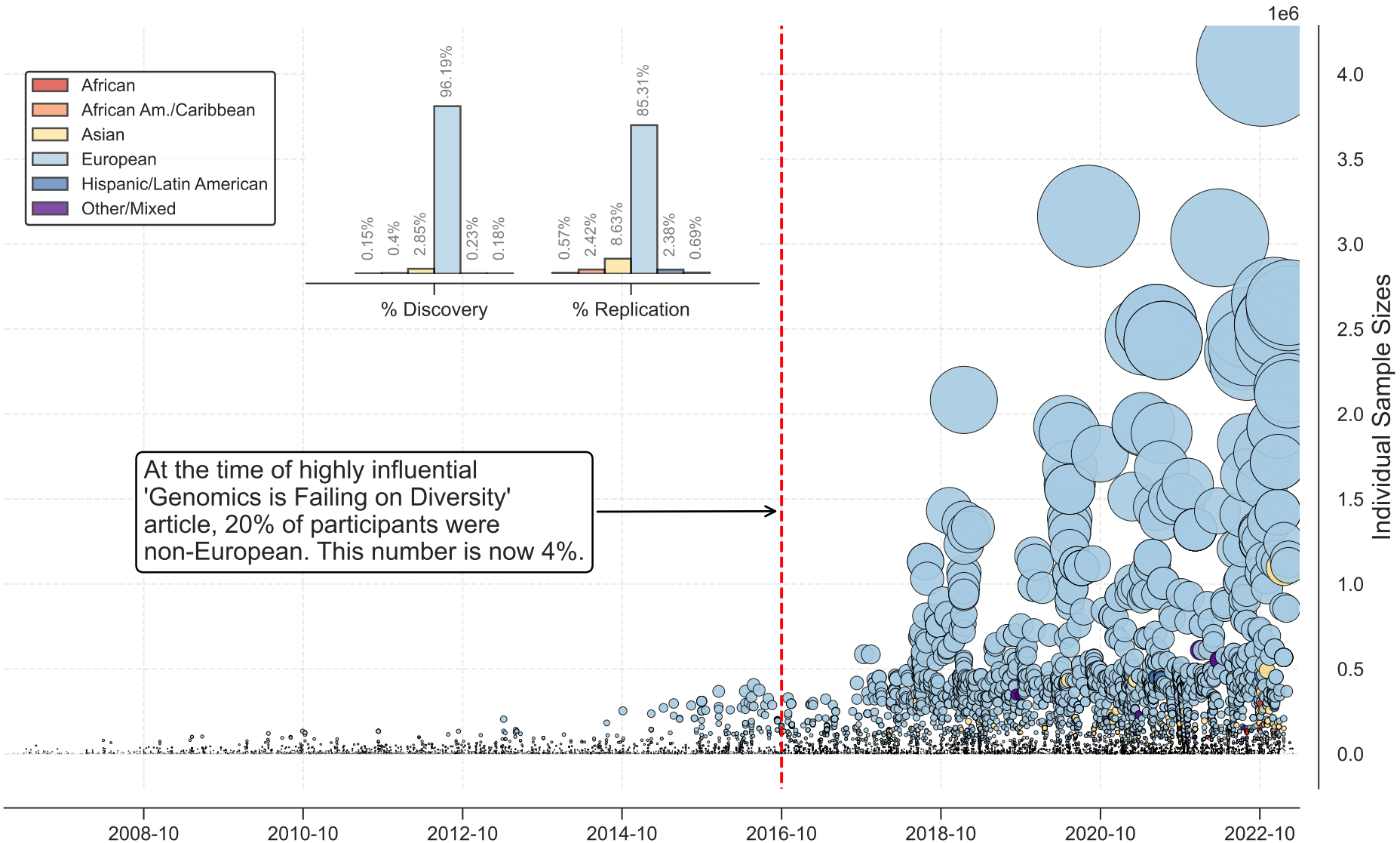
The Problem

- PRS provide predictive power above combined clinical risk factors for many traits in **European** populations.
- However, genetic results discovered in one ancestry are not 'portable':
  - PRS have attenuated cross-population predictive performance.
  - Reductions in portability even in subgroups of European ancestries.
  - Shown in multiple empirical studies, due to genetic architecture, allele frequency, LD patterns, phenotype definition, GxE, etc.
- This is a big problem, because [almost] all large datasets are European.
- If 'discovering' in European ancestry, benefits don't 'replicate' out-of-sample.
- This reduces clinical value in external populations, exacerbating inequalities.
- Public health debate about 'personalized medicine': personalised for who?

The Problem

- PRS provide predictive power above combined clinical risk factors for many traits in **European** populations.
- However, genetic results discovered in one ancestry are not 'portable':
  - PRS have attenuated cross-population predictive performance.
  - Reductions in portability even in subgroups of European ancestries.
  - Shown in multiple empirical studies, due to genetic architecture, allele frequency, LD patterns, phenotype definition, GxE, etc.
- This is a big problem, because [almost] all large datasets are European.
- If 'discovering' in European ancestry, benefits don't 'replicate' out-of-sample.
- This reduces clinical value in external populations, exacerbating inequalities.
- Public health debate about 'personalized medicine': personalised for who?
- Analysis of diverse populations **also** improves discovery for complex traits!

Majority of studies are on people of European Ancestry



This is all made possible by **amazing** data orginators.

GWAS Catalog

PubMed

Europe PubMed Central

Introducing the GWAS Diversity Monitor

- **However**: these continually updating datasets allow us to do more.

Introducing the GWAS Diversity Monitor

- **However**: these continually updating datasets allow us to do more.
- Introducing the GWAS Diversity Monitor: gwasdiversitymonitor.com.

Introducing the GWAS Diversity Monitor

- **However**: these continually updating datasets allow us to do more.
- Introducing the GWAS Diversity Monitor: gwasdiversitymonitor.com.
- Checks GWAS Catalog every day at midnight (cron), and updates.

Introducing the GWAS Diversity Monitor

- **However**: these continually updating datasets allow us to do more.
- Introducing the GWAS Diversity Monitor: gwasdiversitymonitor.com.
- Checks GWAS Catalog every day at midnight (cron), and updates.
- Built on familiar stack of `D3.js` and `Flask`. All code on `GitHub`:

github.com/oxforddemsci/gwasdiversitymonitor

Introducing the GWAS Diversity Monitor

- **However**: these continually updating datasets allow us to do more.
- Introducing the GWAS Diversity Monitor: gwasdiversitymonitor.com.
- Checks GWAS Catalog every day at midnight (cron), and updates.
- Built on familiar stack of `D3.js` and `Flask`. All code on `GitHub`:

  github.com/oxforddemsci/gwasdiversitymonitor

- The app itself runs on an AWS LightSail instance.

Introducing the GWAS Diversity Monitor

- **However**: these continually updating datasets allow us to do more.
- Introducing the GWAS Diversity Monitor: gwasdiversitymonitor.com.
- Checks GWAS Catalog every day at midnight (cron), and updates.
- Built on familiar stack of `D3.js` and `Flask`. All code on `GitHub`:

  github.com/oxforddemsci/gwasdiversitymonitor

- The app itself runs on an AWS LightSail instance.
- Short commentary announced the release of the app in March 2020.

Introducing the GWAS Diversity Monitor

- **However**: these continually updating datasets allow us to do more.
- Introducing the GWAS Diversity Monitor: gwasdiversitymonitor.com.
- Checks GWAS Catalog every day at midnight (cron), and updates.
- Built on familiar stack of `D3.js` and `Flask`. All code on `GitHub`:

  github.com/oxforddemsci/gwasdiversitymonitor

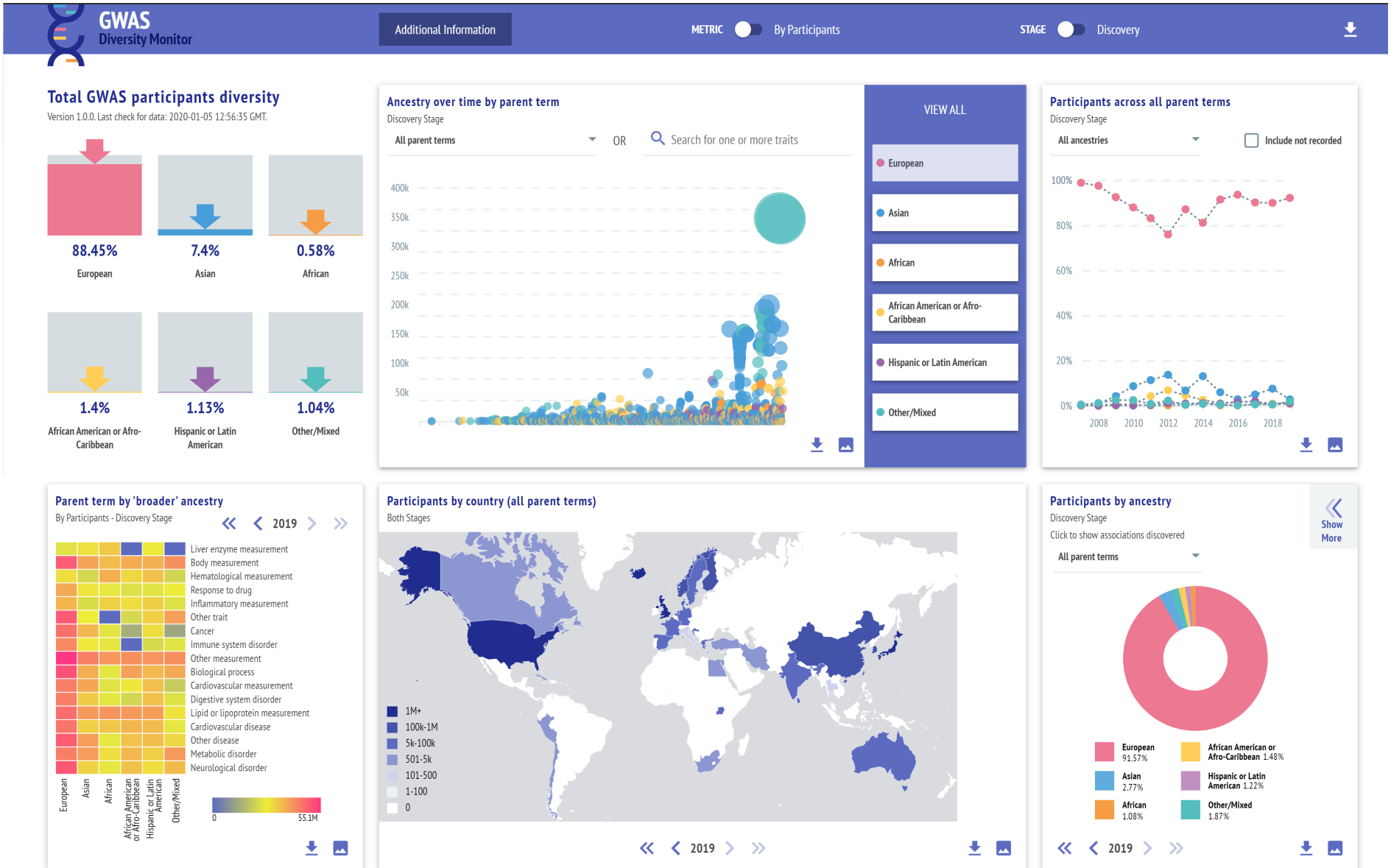
- The app itself runs on an AWS LightSail instance.
- Short commentary announced the release of the app in March 2020.
- It provides the chance for dynamic **accountability** and reflection.

Introducing the GWAS Diversity Monitor

- **However**: these continually updating datasets allow us to do more.
- Introducing the GWAS Diversity Monitor: gwasdiversitymonitor.com.
- Checks GWAS Catalog every day at midnight (cron), and updates.
- Built on familiar stack of `D3.js` and `Flask`. All code on `GitHub`:

  github.com/oxforddemsci/gwasdiversitymonitor

- The app itself runs on an AWS LightSail instance.
- Short commentary announced the release of the app in March 2020.
- It provides the chance for dynamic **accountability** and reflection.
- Used in multiple high profile agencies, such as the NIH.


GWAS
Diversity Monitor
Additional Information
METRIC
By Participants
STAGE
Discovery
Total GWAS participants diversity
Version 1.0.0. Last check for data: 2020-01-05 12:56:35 GMT.
88.45%
European
7.4%
Asian
0.58%
African
1.4%
African American or Afro-Caribbean
1.13%
Hispanic or Latin American
1.04%
Other/Mixed
Ancestry over time by parent term
Discovery Stage
All parent terms
OR
Search for one or more traits
400k
350k
300k
250k
200k
150k
100k
50k
VIEW ALL
European
Asian
African
African American or Afro-Caribbean
Hispanic or Latin American
Other/Mixed
Participants across all parent terms
Discovery Stage
All ancestries
Include not recorded
100%
80%
60%
40%
20%
0%
2008
2010
2012
2014
2016
2018
Parent term by 'broader' ancestry
By Participants - Discovery Stage
2019
Liver enzyme measurement
Body measurement
Hematological measurement
Response to drug
Inflammatory measurement
Other trait
Cancer
Immune system disorder
Other measurement
Biological process
Cardiovascular measurement
Digestive system disorder
Lipid or lipoprotein measurement
Cardiovascular disease
Other disease
Metabolic disorder
Neurological disorder
European
Asian
African
African American or Afro-Caribbean
Hispanic or Latin American
Other/Mixed
0
55.1M
Participants by country (all parent terms)
Both Stages
1M+
100k-1M
5k-100k
501-5k
101-500
1-100
0
2019
Participants by ancestry
Discovery Stage
Click to show associations discovered
Show More
All parent terms
European 91.57%
African American or Afro-Caribbean 1.48%
Asian 2.77%
Hispanic or Latin American 1.22%
African 1.08%
Other/Mixed 1.87%
2019

The Importance of Funders

- Monitoring inclusion necessary but insufficient to enact change.

The Importance of Funders

- Monitoring inclusion necessary but insufficient to enact change.
- Small number of funders enable **large majority** of GWAS research.

The Importance of Funders

- Monitoring inclusion necessary but insufficient to enact change.
- Small number of funders enable **large majority** of GWAS research.
- The NHLBI NIH exclusively funds $\sim$**20%** of *all* GWAS authorships.

The Importance of Funders

- Monitoring inclusion necessary but insufficient to enact change.
- Small number of funders enable **large majority** of GWAS research.
- The NHLBI NIH exclusively funds ∼**20%** of *all* GWAS authorships.
- ∼**83%** of all GWAS funding comes from the US, ∼**15%** from UK.

The Importance of Funders

- Monitoring inclusion necessary but insufficient to enact change.
- Small number of funders enable **large majority** of GWAS research.
- The NHLBI NIH exclusively funds ∼**20%** of *all* GWAS authorships.
- ∼**83%** of all GWAS funding comes from the US, ∼**15%** from UK.
- Dominent UK funders: MRC (∼**9.3%**), WT (∼**3.9%**), CRUK (∼**1.2%**).

The Importance of Funders

- Monitoring inclusion necessary but insufficient to enact change.
- Small number of funders enable **large majority** of GWAS research.
- The NHLBI NIH exclusively funds ∼**20%** of *all* GWAS authorships.
- ∼**83%** of all GWAS funding comes from the US, ∼**15%** from UK.
- Dominent UK funders: MRC (∼**9.3%**), WT (∼**3.9%**), CRUK (∼**1.2%**).
- **Q:** What is the most frequently observed GrantID returned from PubMed?

The Importance of Funders

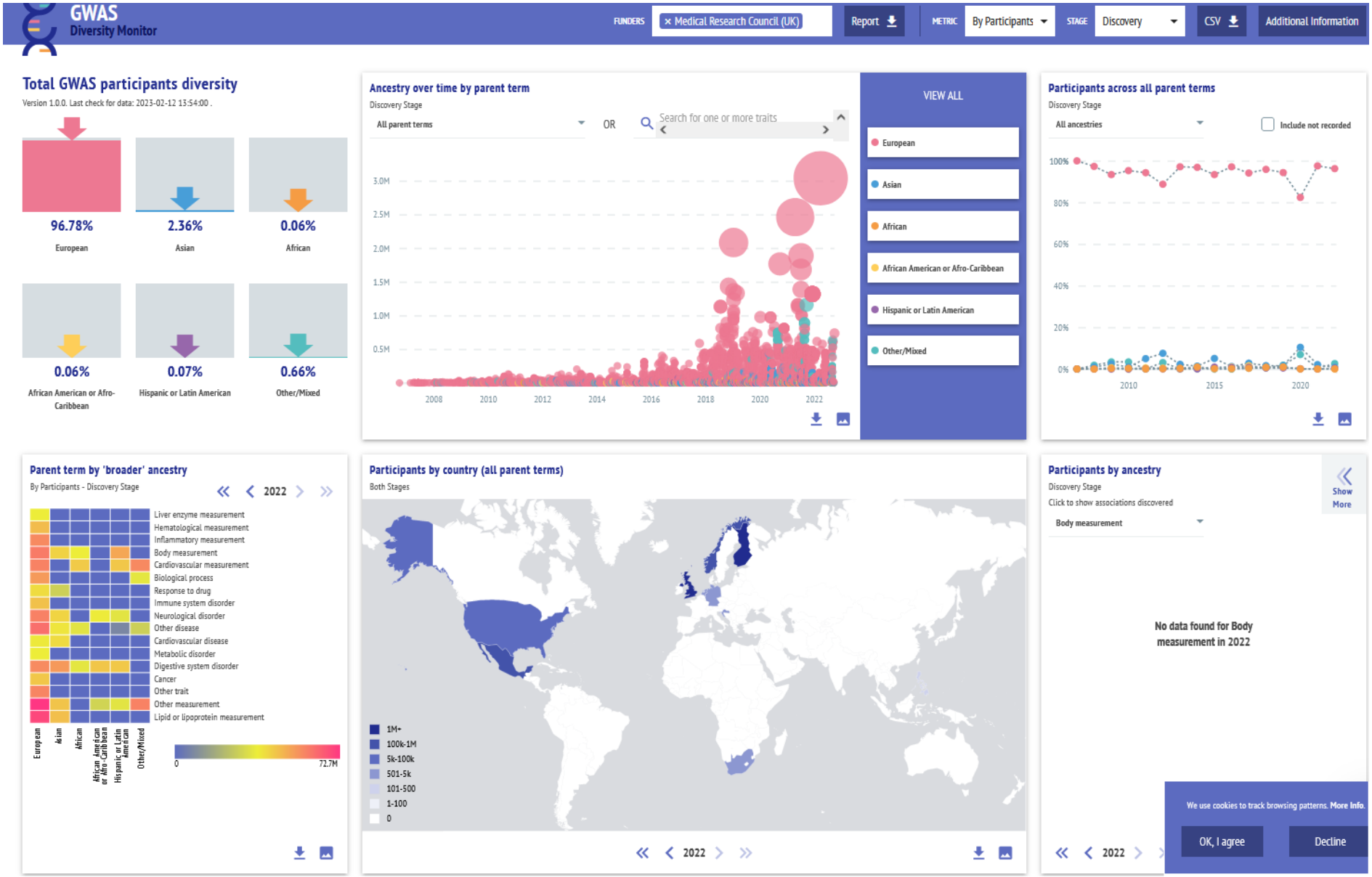
- Monitoring inclusion necessary but insufficient to enact change.
- Small number of funders enable **large majority** of GWAS research.
- The NHLBI NIH exclusively funds ∼**20%** of *all* GWAS authorships.
- ∼**83%** of all GWAS funding comes from the US, ∼**15%** from UK.
- Dominent UK funders: MRC (∼**9.3%**), WT (∼**3.9%**), CRUK (∼**1.2%**).
- **Q:** What is the most frequently observed GrantID returned from PubMed?
- **A.** £66m to **MC_qA137853** from the MRC, beginning in April 2006...
    - The stratified nature of the **UK Biobank** is very well documented.

The Importance of Funders

- Monitoring inclusion necessary but insufficient to enact change.
- Small number of funders enable **large majority** of GWAS research.
- The NHLBI NIH exclusively funds ∼**20%** of *all* GWAS authorships.
- ∼**83%** of all GWAS funding comes from the US, ∼**15%** from UK.
- Dominent UK funders: MRC (∼**9.3%**), WT (∼**3.9%**), CRUK (∼**1.2%**).
- **Q:** What is the most frequently observed GrantID returned from PubMed?
- **A.** £66m to **MC_qA137853** from the MRC, beginning in April 2006...
  - The stratified nature of the **UK Biobank** is very well documented.
- Highlights importance of the funding (policies) of large genomic datasets.

The Importance of Funders

- Monitoring inclusion necessary but insufficient to enact change.
- Small number of funders enable **large majority** of GWAS research.
- The NHLBI NIH exclusively funds ∼**20%** of *all* GWAS authorships.
- ∼**83%** of all GWAS funding comes from the US, ∼**15%** from UK.
- Dominent UK funders: MRC (∼**9.3%**), WT (∼**3.9%**), CRUK (∼**1.2%**).
- **Q:** What is the most frequently observed GrantID returned from PubMed?
- **A.** £66m to **MC_qA137853** from the MRC, beginning in April 2006...
  - The stratified nature of the **UK Biobank** is very well documented.
- Highlights importance of the funding (policies) of large genomic datasets.
- 'All of Us' sets ground-breaking strategy regarding Diversity and Inclusion.


GWAS Diversity Monitor
FUNDERS
× Medical Research Council (UK)
Report
METRIC
By Participants
STAGE
Discovery
CSV
Additional Information
Total GWAS participants diversity
Version 1.0.0. Last check for data: 2023-02-12 13:54:00 .
96.78%
European
2.36%
Asian
0.06%
African
0.06%
African American or Afro-Caribbean
0.07%
Hispanic or Latin American
0.66%
Other/Mixed
Ancestry over time by parent term
Discovery Stage
All parent terms
OR
Search for one or more traits
3.0M
2.5M
2.0M
1.5M
1.0M
0.5M
2008
2010
2012
2014
2016
2018
2020
2022
VIEW ALL
European
Asian
African
African American or Afro-Caribbean
Hispanic or Latin American
Other/Mixed
Participants across all parent terms
Discovery Stage
All ancestries
Include not recorded
100%
80%
60%
40%
20%
0%
2010
2015
2020
Parent term by 'broader' ancestry
By Participants - Discovery Stage
2022
Liver enzyme measurement
Hematological measurement
Inflammatory measurement
Body measurement
Cardiovascular measurement
Biological process
Response to drug
Immune system disorder
Neurological disorder
Other disease
Cardiovascular disease
Metabolic disorder
Digestive system disorder
Cancer
Other trait
Other measurement
Lipid or lipoprotein measurement
European
Asian
African
African American or Afro-Caribbean
Hispanic or Latin American
Other/Mixed
0
72.7M
Participants by country (all parent terms)
Both Stages
1M+
100k-1M
5k-100k
501-5k
101-500
1-100
0
2022
Participants by ancestry
Discovery Stage
Click to show associations discovered
Body measurement
Show More
No data found for Body measurement in 2022
2022
We use cookies to track browsing patterns. More Info.
OK, I agree
Decline

Caveats/Limitations/Next Steps

There are necessarily a number of limitations (and other solutions):

Caveats/Limitations/Next Steps

There are necessarily a number of limitations (and other solutions):

- We mostly rely on **ancestry groupings** in the NHGRI-EBI Catalog:
  - The ancestries are often self-reported.
  - These classifications fail to account for genetic admixture.
  - Many new tools account for admixture/cross-population prediction.

Caveats/Limitations/Next Steps

There are necessarily a number of limitations (and other solutions):

- We mostly rely on **ancestry groupings** in the NHGRI-EBI Catalog:
  - The ancestries are often self-reported.
  - These classifications fail to account for genetic admixture.
  - Many new tools account for admixture/cross-population prediction.
- In terms of purely **technical concerns**:
  - The app is currently slow: we've just put up a `CONTRIBUTING.md`!
  - Yet to determine content of automatically (funder) generated reports.

Caveats/Limitations/Next Steps

There are necessarily a number of limitations (and other solutions):

- We mostly rely on **ancestry groupings** in the NHGRI-EBI Catalog:
  - The ancestries are often self-reported.
  - These classifications fail to account for genetic admixture.
  - Many new tools account for admixture/cross-population prediction.
- In terms of purely **technical concerns**:
  - The app is currently slow: we've just put up a `CONTRIBUTING.md`!
  - Yet to determine content of automatically (funder) generated reports.
- What does the post-GWAS era look like?

Caveats/Limitations/Next Steps

There are necessarily a number of limitations (and other solutions):

- We mostly rely on **ancestry groupings** in the NHGRI-EBI Catalog:
  - The ancestries are often self-reported.
  - These classifications fail to account for genetic admixture.
  - Many new tools account for admixture/cross-population prediction.
- In terms of purely **technical concerns**:
  - The app is currently slow: we've just put up a `CONTRIBUTING.md`!
  - Yet to determine content of automatically (funder) generated reports.
- What does the post-GWAS era look like?
- Please do get in contact with all and any suggestions for improvements!

contact@gwasdiversitymonitor.com

Thanks to all the Funders and (much) Broader Team!




European Research Council
Established by the European Commission

LEVERHULME TRUST

UNIVERSITY OF OXFORD

The British Academy