

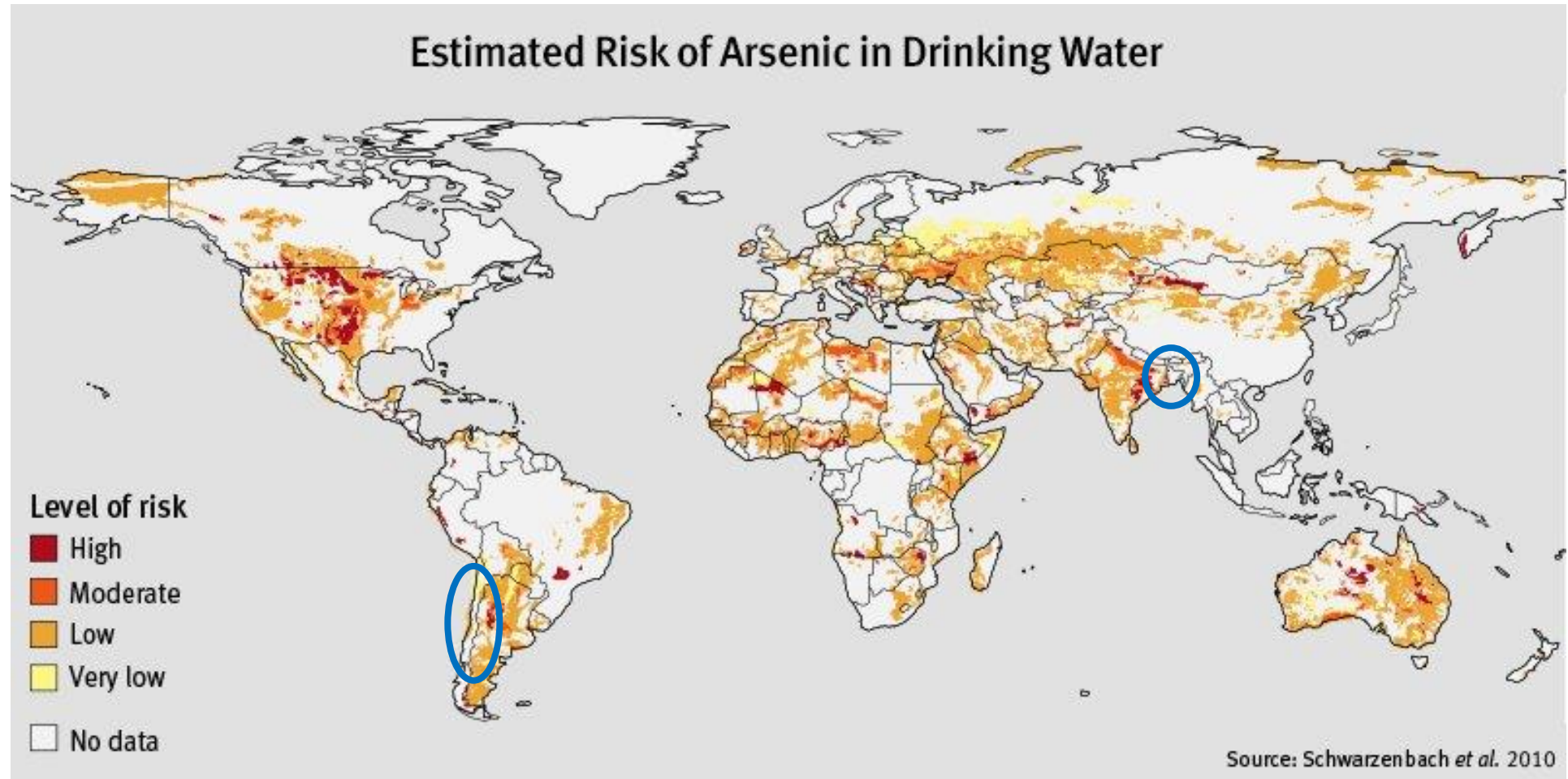
# Arsenic Epigenetics **META:** **Meta-analysis of Epigenome Data on Arsenic**

Risk e-Learning Webinar Series, August 3, 2021

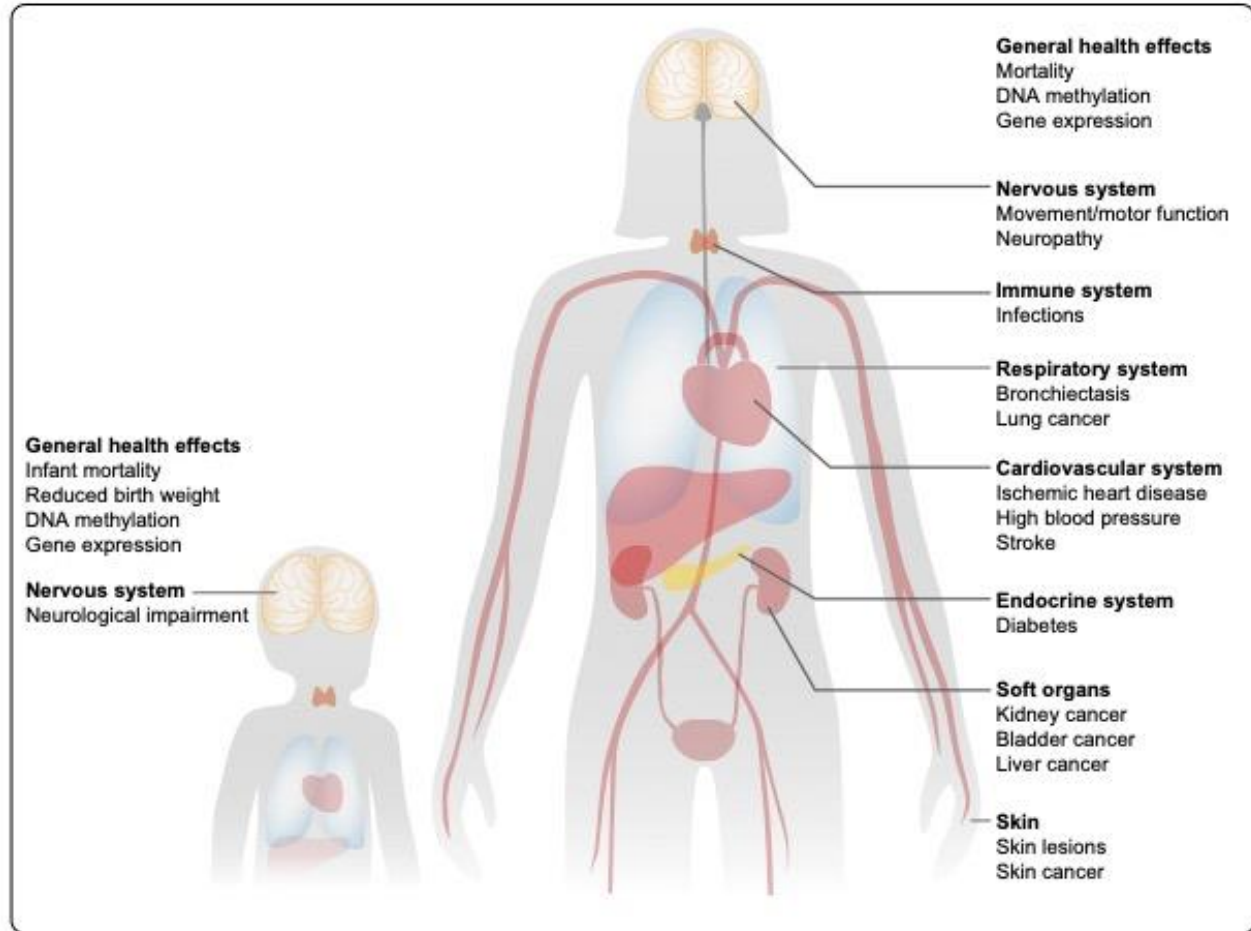
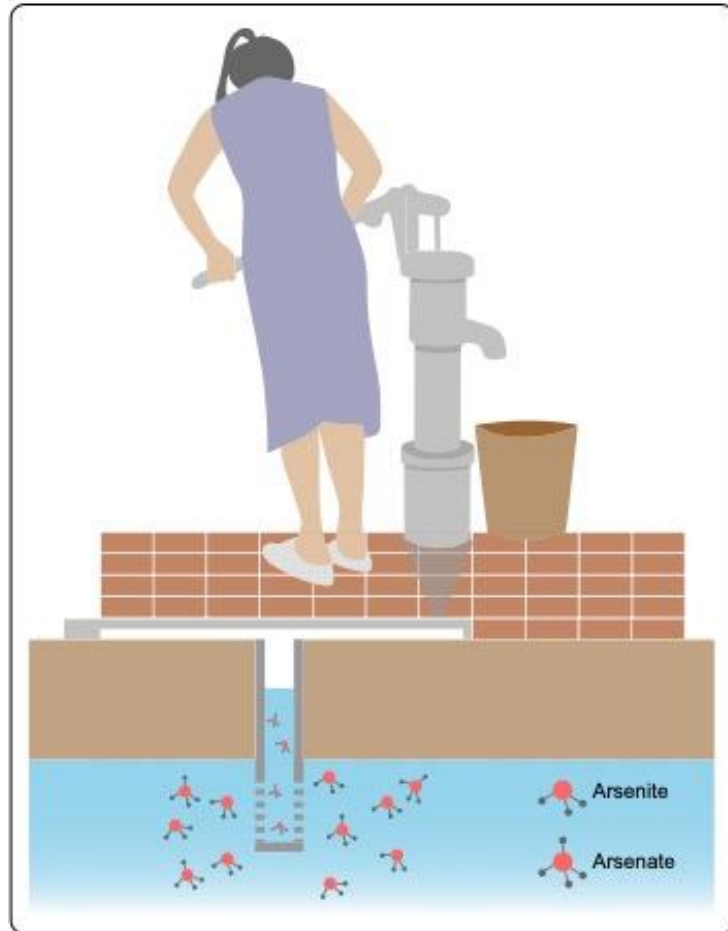
Anne Bozack, MPH, PhD

Andres Cardenas, MPH, PhD

# Arsenic exposure and health effects

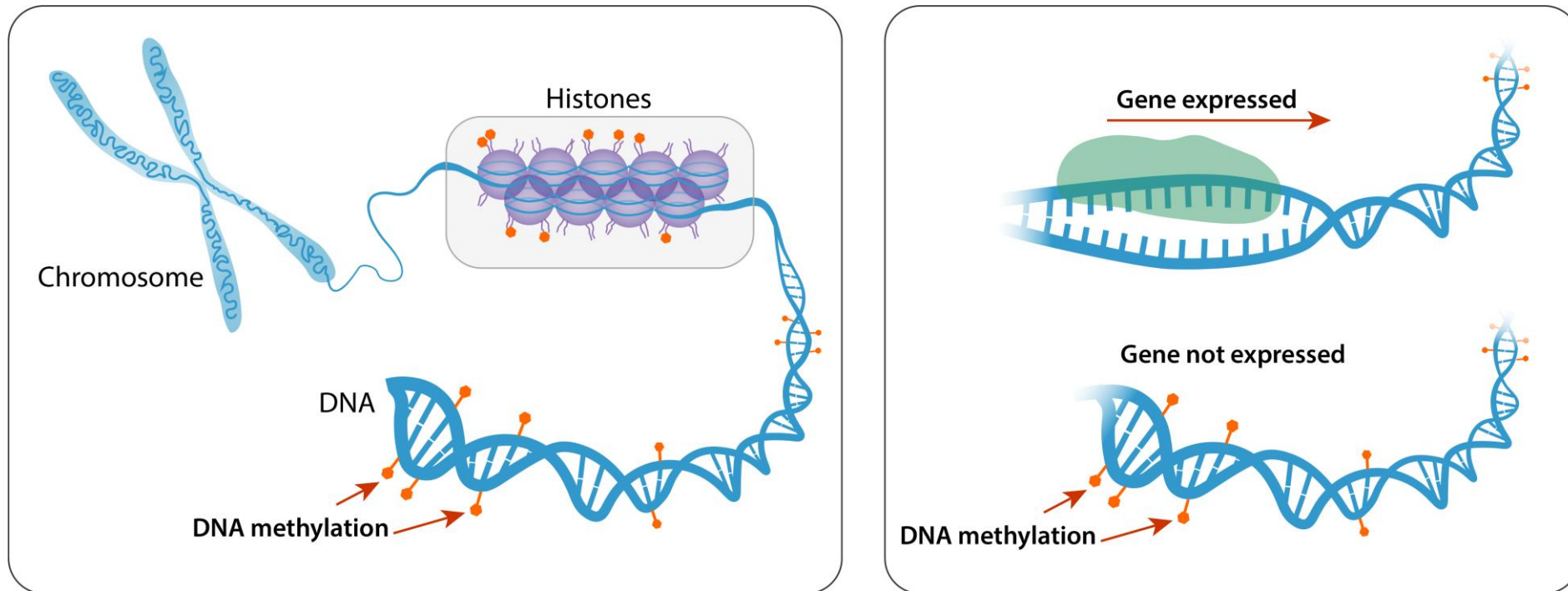


# Arsenic exposure and health effects



# Arsenic exposure and health effects

- Arsenic-related health risks persist after exposure has ended.
  - Epigenetic dysregulation may be a mechanistic link between As and health outcomes.



# General overview

- **Summary**

- Leverage previously measured Epigenome-Wide DNA methylation data across SRP centers for a meta-analysis of arsenic exposure on the epigenome of human cohorts
  - Addresses the question as to whether epigenetic biomarkers of As exposure are generalizable
- Goal is to develop a framework, protocols, open-source code, and associated workflow that can be utilized to meta-analyze multiple EWAS related to environmental exposures (Epigenetics Consortium of Environmental Exposures)

|                           | University of California, Berkeley                           | Columbia University   |
|---------------------------|--|---|
| <b>Point of contact</b>   | Andres Cardenas  | Mary Gamble   |
| <b>Lead project title</b> | Exposomics and Arsenic Epidemiology                          | Impact of Nutrition on Arsenic-Induced Epigenetic Dysregulation |
| <b>Other partners</b>     | Craig Steinmaus; Martyn Smith; Waverly Wei, Philippe Boileau | Ana Navas-Acien; <b>Anne Bozack</b>                             |

# Inputs and actions

## Inputs

- **Existing data sets**: Columbia SRP DNA methylation data from Bangladeshi adults exposed to arsenic (**urinary and water**); UC Berkeley cohort from Northern Chile of adults exposed to arsenic early in life (**prenatal vs post**)
- **Variables**: High dimensional DNA methylation data (**450K** or **850K CpG** sites in the human genome); historical As exposure and biomarkers, and demographic characteristics
- **Repositories**: Data currently stored locally at each SRP center but not systematically preserved/annotated

## Actions: how are we achieving F, A, I, and or R

- Analytical code is **findable** internally/**externally** by users at each center by navigating a well-annotated GitHub repository ([https://github.com/annebozack/SRP\\_arsenic\\_DNA\\_methylation\\_metaanalysis](https://github.com/annebozack/SRP_arsenic_DNA_methylation_metaanalysis))
- Summary results **accessible** by sharing our analytical protocol and code : [https://github.com/annebozack/SRP\\_arsenic\\_DNA\\_methylation\\_metaanalysis](https://github.com/annebozack/SRP_arsenic_DNA_methylation_metaanalysis)
- We will increase **interoperability** as summary EWAS findings can be integrated with other omics results (OSF)
- By preserving our data and annotated code we will ensure data is **reusable** for trainees and investigators. (**Epigenetic Aging Biomarkers**)

# Collaboration tools

## GitHub

- Created a shared repository to collaborate on development of data processing and analysis pipeline
- Ensured that collaborators had access to the most recent code versions
- Repository made publicly available for other researchers to access data processing and analysis pipeline



## Box

- Used to securely store/transfer EWAS results between centers
- Convenient upload/download of large datasets (e.g., output from ~450,000 and 850,00 analyses)



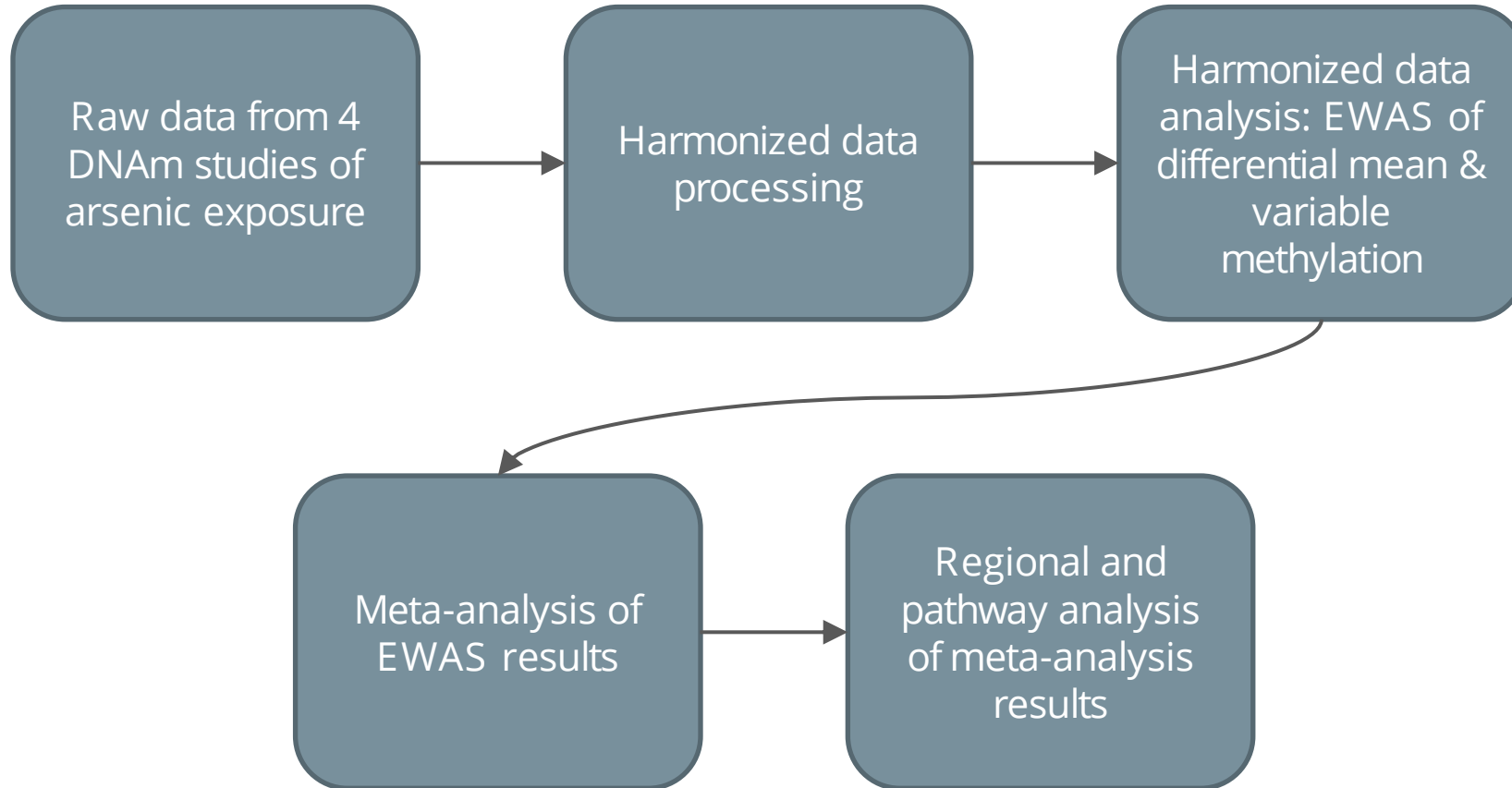
## Google docs

- Allowed for collaboration and version control during manuscript preparation



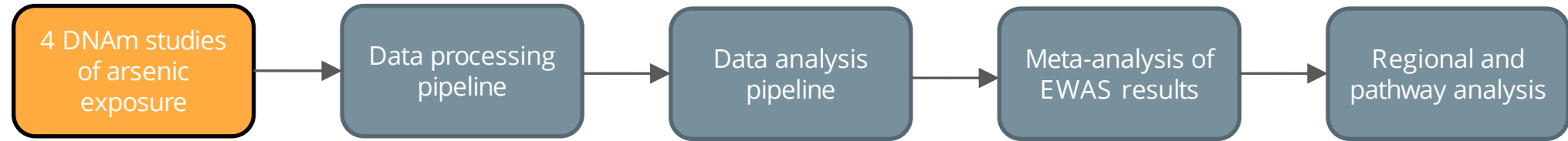
Google Docs

# Actions: Data processing and analysis pipeline





# Actions: Data processing and analysis pipeline



- Two study locations: Bangladesh and Chile
- Chile: two different tissues (buccal and blood cells)
- Bangladesh: two epigenomics platforms: 450K and EPIC (850K)
- Comparison of adult chronic exposure (Bangladesh) vs high fetal exposure (Chile)

|                                      | Chile, PBMCs<br>(N = 40) <sup>a</sup> |       | Chile, buccal cells<br>(N = 39) <sup>a</sup> |       | Bangladesh, 450K<br>(N = 48) |        | Bangladesh, 850K<br>(N = 32) |        |
|--------------------------------------|---------------------------------------|-------|--|-------|------------------------------|--------|------------------------------|--------|
|                                      | n                                     | %     | n  | %     | n                            | %      | n                            | %      |
| Age, years, mean (SD)                | 48.7                                  | (4.7) | 48.7   | (4.7) | 39.7                         | (8.1)  | 41.7                         | (6.3)  |
| Male                                 | 21                                    | 52.5  | 20   | 51.3  | 48                           | 100.0% | 32                           | 100.0% |
| Ever smoker                          | 16                                    | 40.0  | 16   | 41.0  | 21                           | 43.8%  | 20                           | 62.5%  |
| Prenatal/early life arsenic exposure | 20                                    | 50.0  | 19   | 48.7  | -                            | -      | -                            | -      |
| High arsenic exposure <sup>b</sup>   | -                                     | -     | -  | -     | 23                           | 47.9%  | 11                           | 34.4%  |

a. 850K; PBMC and buccal cell samples from the same study participants. b.  $\geq 100$   $\mu\text{g/L}$  water arsenic for 450K analyses and 104  $\mu\text{g/L}$  water arsenic for 850K analyses.

# Actions: Data processing and analysis pipeline



annebozack / SRP\_arsenic\_DNAm\_metaanalysis

Code Issues Pull requests Actions Projects Wiki Security Insights

main 1 branch 0 tags

Go to file Add file Code

annebozack Add repository files 628584f on Jan 9 3 commits

- DMR-meta-analysis Add repository files last month
- EWAS-results Add repository files last month
- bangladesh-study Add repository files last month
- chile-study Add repository files last month
- helper-scripts Add repository files last month
- LICENSE Add repository files last month
- README.md Add repository files last month
- Superfund-Methylation.Rproj Add repository files last month

README.md

### Exposure to arsenic at different life-stages and DNA methylation meta-analysis in buccal cells and leukocytes

This repository contains the necessary scripts to reproduce the analysis of Bozack et al.'s "Exposure to arsenic at different life-stages and DNA methylation meta-analysis in buccal cells and leukocytes". A preprint of the manuscript can be found here (insert link).

The organization of the repository is as follows:

- The `bangladesh-study` folder contains the scripts and results associated with the Bangladesh DNAm studies.
- The `chile-study` directory is made up of subdirectories containing the code and results associated with the buccal cell and the PBMC DNAm analyses. It also holds a directory with notebooks assessing the within-participant buccal-PBMC sample similarities, and a directory with a notebook of descriptive statistics used to create Table 1 in the accompanying paper.
- The `DMR-meta-analysis` folder contains the `DMR-meta-analysis.Rmd` notebook, which details the meta-analysis procedure and summarized the results output by `comb-p`.
- The `helper-scripts` directory contains multiple helper files used to perform the analyses described in the paper.

## Bangladesh: Leukocytes

Load Data

Internal Quality Control

Outlier information

Pre Filtering

**Pre Filtering Density**

Pre Filtering Beta Values

Pre Filtering Dendrogram

Pre-Filtering SVD

Post Filtering

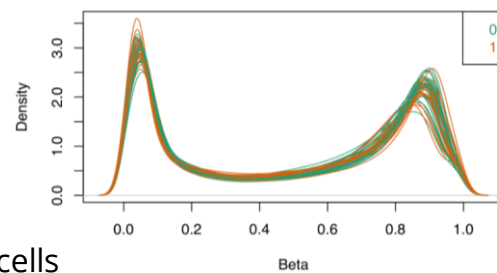
Normalization

Batch correction

### Pre Filtering Density

```
knitrr::include_graphics("QC/QC_prefiltering/raw_densityplot.png")
```

### Density plot of raw data (485512 probes)



Required Packages

DMPs

DMP datasets for common CpGs

Run in METAL

Without cell type adjustment

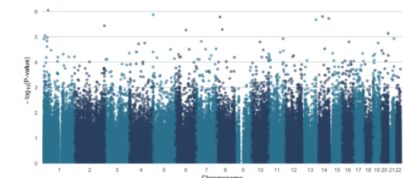
**Blood only**

Including buccal

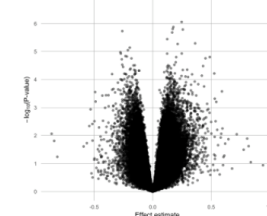
With cell type adjustment

DVPs

```
knitrr::include_graphics("manhattan_metal_blood_noCell.png")
```



```
knitrr::include_graphics("volcano_metal_blood_noCell.png")
```



### Top 100 probes

|        | Name       | Effect  | StdErr | Pvalue    | Direction | PFDR      | PBonf     | chr   | pos       | UCSC_RefGene_N   |
|--------|------------|---------|--------|-----------|-----------|-----------|-----------|-------|-----------|------------------|
| 93     | cg00004257 | 0.2473  | 0.0503 | 0.0000009 | +++       | 0.1315697 | 0.3325217 | chr1  | 34629175  | CSMD2            |
| 316516 | cg22938488 | 0.1765  | 0.0365 | 0.0000013 | ++        | 0.1315697 | 0.5045183 | chr5  | 1501670   | LPCAT1           |
| 114376 | cg07490485 | 0.1711  | 0.0356 | 0.0000016 | +++       | 0.1315697 | 0.5924411 | chr14 | 50432614  |                  |
| 194923 | cg13490635 | 0.2572  | 0.0537 | 0.0000016 | +++       | 0.1315697 | 0.6207424 | chr8  | 30242021  | RBPM5, RBPM5, RE |
| 17421  | cg01050273 | -0.2598 | 0.0545 | 0.0000019 | ---       | 0.1315697 | 0.7064011 | chr14 | 101152705 |                  |
| 129233 | cg08528486 | 0.1963  | 0.0350 | 0.0000021 | +-        | 0.1315697 | 0.7894183 | chr13 | 113648767 | MCF2L, MCF2L     |
| 248565 | cg17342864 | 0.1239  | 0.0268 | 0.0000036 | +++       | 0.1950905 | 1.0000000 | chr2  | 239973967 | HDACA            |
| 39707  | cg02463029 | 0.3626  | 0.0795 | 0.0000051 | +++       | 0.2263687 | 1.0000000 | chr8  | 48297271  | KJAA0146         |
| 359036 | cg26299756 | 0.2130  | 0.0468 | 0.0000054 | +++       | 0.2263687 | 1.0000000 | chr6  | 85478807  |                  |
| 73319  | cg04685632 | -0.1960 | 0.0437 | 0.0000073 | --        | 0.2749002 | 1.0000000 | chr20 | 62082611  | KCNQ2, KCNQ2, KC |
| 182880 | cg12547807 | -0.2121 | 0.0478 | 0.0000090 | --        | 0.2996167 | 1.0000000 | chr1  | 9473751   |                  |
| 46132  | cg02892755 | -0.2760 | 0.0627 | 0.0001106 | --        | 0.2996167 | 1.0000000 | chr1  | 28103358  | STX12            |
| 233149 | cg16190478 | 0.2482  | 0.0566 | 0.0001118 | +++       | 0.2996167 | 1.0000000 | chr21 | 45789122  | TRPM2            |

## Chile: Buccal cells

Required Packages

Load Data

Internal Quality Control

Outlier information

Pre Filtering

**Pre Filtering Density**

Pre Filtering Beta Values

Pre Filtering Dendrogram

Pre-Filtering SVD

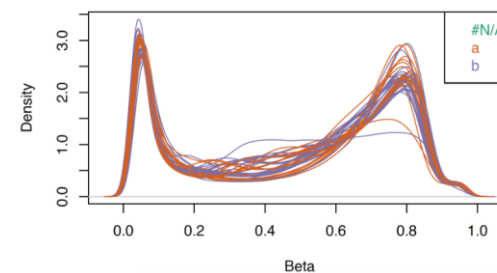
Post Filtering

Normalization

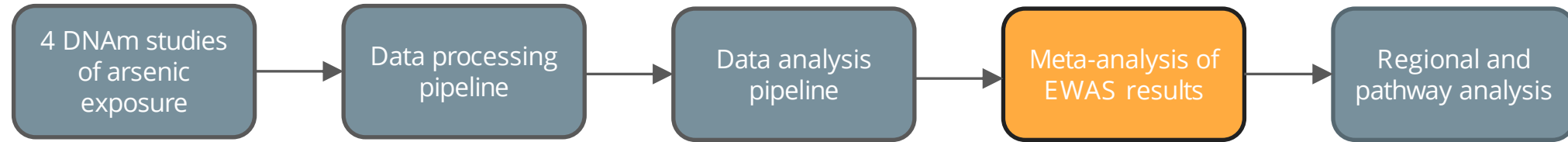
### Pre Filtering Density

```
knitrr::include_graphics("pre-density.png")
```

### Density plot of raw data (866836 probes)



# Actions: Data processing and analysis pipeline



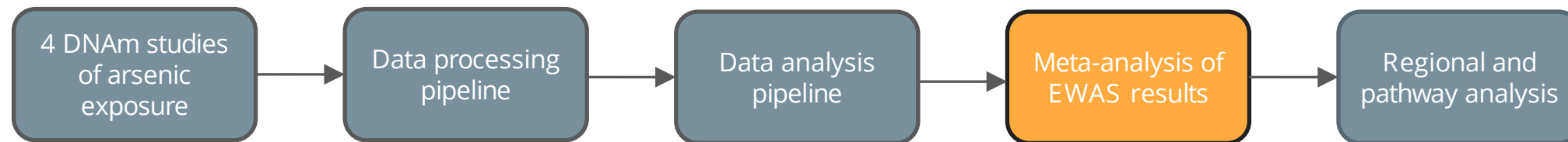
## Summary of results of individual EWAS

|                     | Common probes <sup>a</sup> |                      |
|---------------------|----------------------------|----------------------|
| <b>DMPs</b>         | <b>p &lt; 0.05</b>         |                      |
| Chile, PBMCs        | 23,116                     |                      |
| Chile, buccal cells | 21,336                     |                      |
| Bangladesh, 450K    | 18,301                     |                      |
| Bangladesh, 850K    | 7,954                      |                      |
| <b>DVPs</b>         | <b>p &lt; 0.05</b>         | <b>FDR &lt; 0.05</b> |
| Chile, PBMCs        | 23,487                     | 3                    |
| Chile buccal, cells | 20,735                     | 4                    |
| Bangladesh, 450K    | 16,904                     | 2                    |
| Bangladesh, 850K    | 26,155                     | 24                   |

DMP: differentially methylated position; DVP: differentially variable position. Adjusted for cell type proportions, age, and smoking status. a. 377,351 included in all four EWAS.

No DMPs at FDR < 0.05 identified in individual EWAS.

# Actions: Data processing and analysis pipeline



## Summary of results of individual EWAS

|                     | Common probes <sup>a</sup> |                      |
|---------------------|----------------------------|----------------------|
| <b>DMPs</b>         | <b>p &lt; 0.05</b>         |                      |
| Chile, PBMCs        | 23,116                     |                      |
| Chile, buccal cells | 21,336                     |                      |
| Bangladesh, 450K    | 18,301                     |                      |
| Bangladesh, 850K    | 7,954                      |                      |
| <b>DVPs</b>         | <b>p &lt; 0.05</b>         | <b>FDR &lt; 0.05</b> |
| Chile, PBMCs        | 23,487                     | 3                    |
| Chile buccal, cells | 20,735                     | 4                    |
| Bangladesh, 450K    | 16,904                     | 2                    |
| Bangladesh, 850K    | 26,155                     | 24                   |

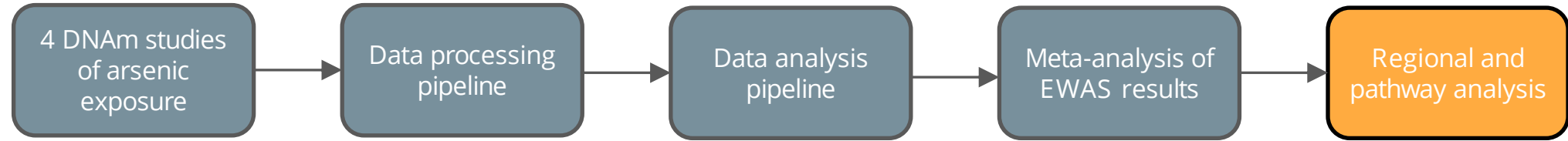
DMP: differentially methylated position; DVP: differentially variable position. Adjusted for cell type proportions, age, and smoking status. a. 377,351 included in all four EWAS.

## Summary of results of meta-analyses

|                      | <b>p &lt; 0.05</b> | <b>FDR &lt; 0.05</b> | <b>λ</b> |
|----------------------|--------------------|----------------------|----------|
| <b>DMPs</b>          |                    |                      |          |
| PBMCs                | 23,361             | 1                    | 1.07     |
| PBMCs + buccal cells | 22,612             | 3                    | 1.06     |
| <b>DVPs</b>          |                    |                      |          |
| PBMCs                | 28,578             | 23                   | 1.17     |
| PBMCs + buccal cells | 28,399             | 19                   | 1.18     |

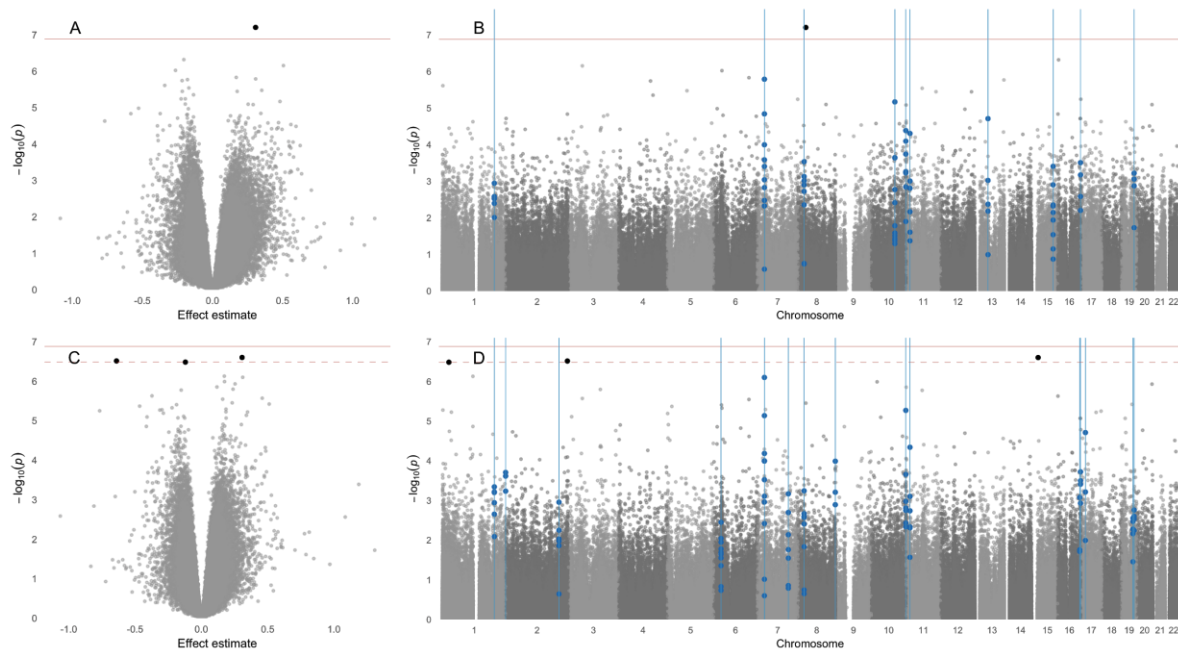
Adjusted for cell type proportions, age, and smoking status. DMP: differentially methylated position; DVP: differentially variable position.

# Actions: Data processing and analysis pipeline



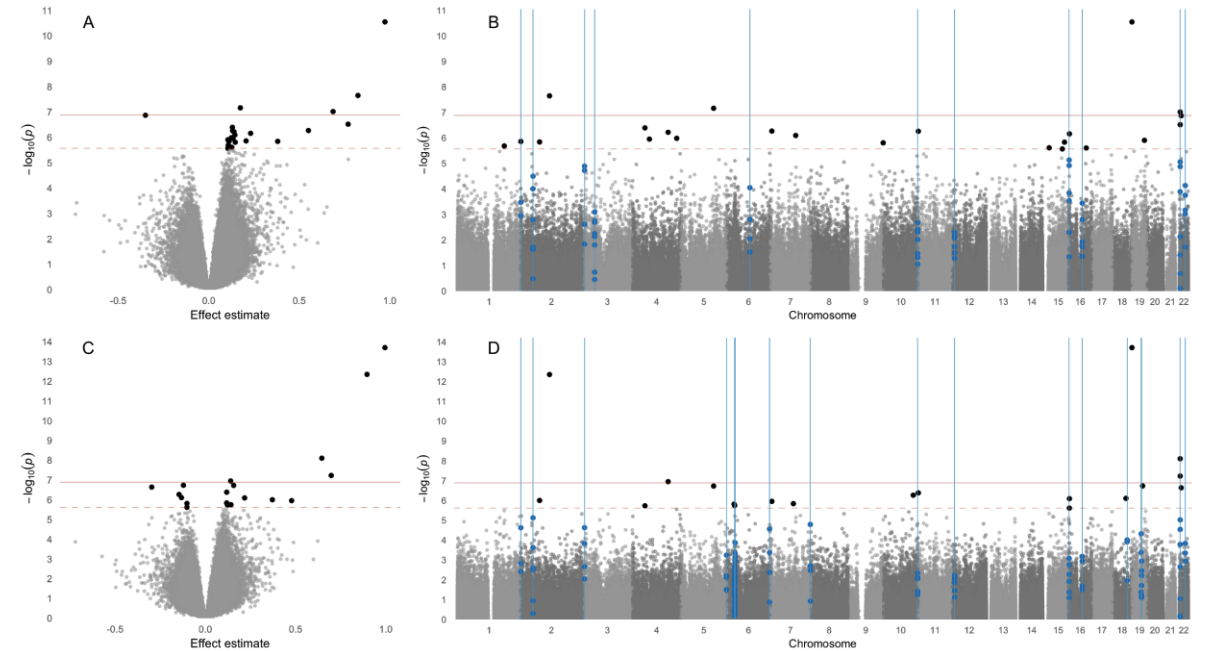
## Differential mean methylation

Top row: PBMC EWAS; Bottom row: PBMC + buccal cell EWAS

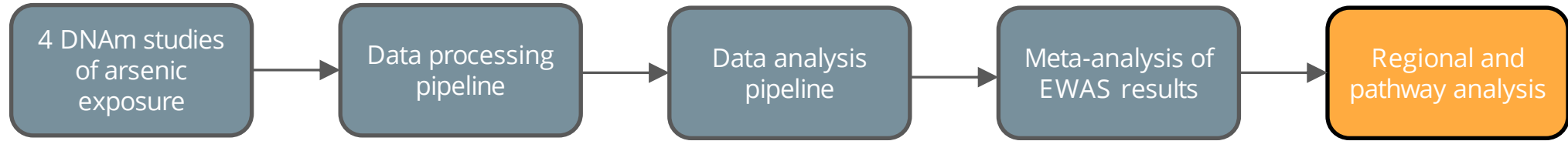


## Differential variability in methylation

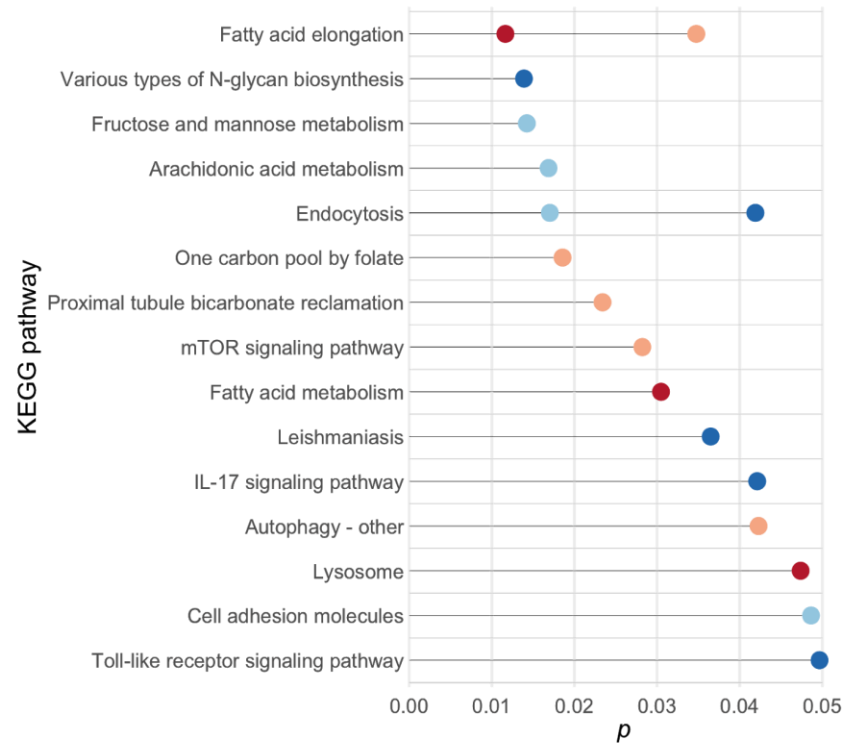
Top row: PBMC EWAS; Bottom row: PBMC + buccal cell EWAS



# Actions: Data processing and analysis pipeline

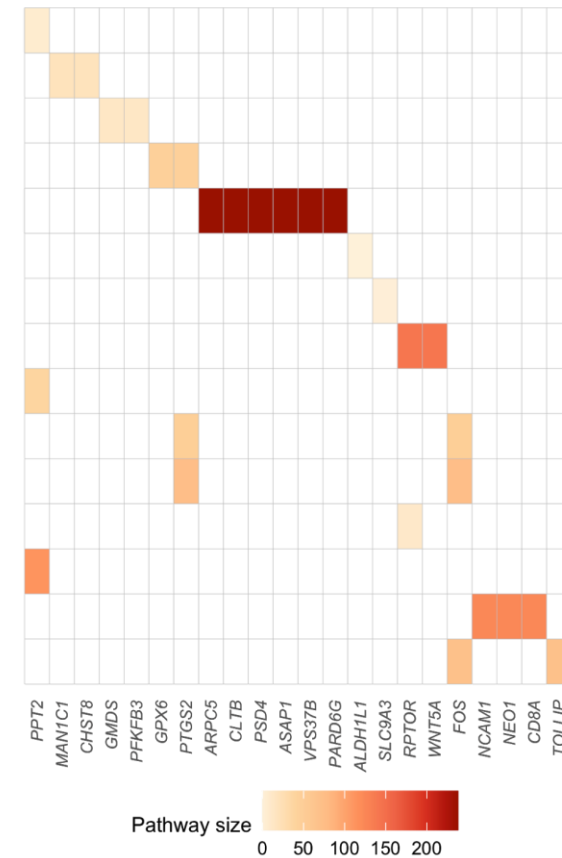


## KEGG pathway analyses



Meta-analysis

- DMPs, PBMCs
- DVPs, PBMCs
- DMPs, PBMCs + buccal cells
- DVPs, PBMCs + buccal cells



Pathway size

0 50 100 150 200

# Actions

**Platforms:** Data processing/analysis pipelines and results available on GitHub; will transfer to Open Science Framework

- GitHub repository: [https://github.com/annebozack/SRP\\_arsenic\\_DNA\\_m\\_metaanalysis](https://github.com/annebozack/SRP_arsenic_DNA_m_metaanalysis)

**Integrating datasets:** Established consistent classification of exposure across datasets; epigenetic measurements and QC

**Communication:** In-person project planning meeting; virtual symposium; weekly virtual meetings

- Virtual symposium: <https://www.youtube.com/watch?v=J3-my0AVIU0>
- GitHub repository to collaborate on developing code
- Google docs to work on manuscript

**Collaborations:** Established ongoing collaboration between UC Berkeley and Columbia SRPs around arsenic-induced epigenetic dysregulation

# Outcomes and deliverables

## Short-term

- Analytical approach for conducting meta-analyses of EWAS across different populations, platforms, and exposures

## Intermediate

- Harmonized data processing and analysis pipeline
- Repository for code and results
- Virtual metal epigenetics symposium: <https://www.youtube.com/watch?v=J3-myoAVIU0>

## Long-Term

- Manuscript describing EWAS meta-analysis approach and findings (*Environ Health*. 2021 Jul 9; 20(1): 79. doi: [10.1186/s12940-021-00754-7](https://doi.org/10.1186/s12940-021-00754-7))
- Code and summary results publicly available
- Possible collaborations with other groups with arsenic and epigenomic data
- Creation of an Environmental Epigenetic Consortium (future)
- Collaboration between Biostatistics students and EHS scientist



# Lessons learned

- Collaboration is key (multiple stakeholders), and reuse of data improves data FAIRness
- Standard QC practices helped us compare data directly
- Improved data curation practices, annotation and storage
- Long-term storage of data with detail information will facilitate reuse
- Center specific analyses allows for equal partnership and shared governance

# Advantages of collaboration and data sharing

- **Scientific question:** reproducibility of arsenic associated epigenetic dysregulation?
  - Pooling data enabled us to increase statistical power
  - Improved generalizability of findings
  - Meta-analyses can yield robust human epigenetic biomarkers
- Two cohorts and multiple tissues improved interpretability of epigenetic signature
- Results differed (*i.e.* cohort specific signals vs. common epigenetic signatures)
- Including more studies could address chronic vs acute exposure signatures
- Future questions that remain are *i)* chronic vs acute As epigenetic signature *ii)* reliability of arsenic exposure biomarker *iii)* expanding to other cohorts

# Recommendations

- **Training**
  - Increasing data FAIRness for all research projects (PIs and trainees)
  - Application of data science methods to existing problems
- **What future activities** are needed to ensure success?
  - Provide incentives for collaborations (i.e., supplemental funds)
  - Increase participation of statisticians and bioinformaticians within and across centers
  - Increase activities/training among statisticians/data scientist and lab scientists
- **What future activities are** needed to foster and advance data sharing?
  - Provide incentives for collaborations (i.e., supplemental funds)
  - Increase participation of statisticians and bioinformaticians within and across centers



**UC BERKELEY**  
**SUPERFUND**  
RESEARCH PROGRAM  
SCIENCE FOR A SAFER WORLD

# Questions?



**SRP**  
Globally Reducing  
Arsenic Exposure

Columbia  
Superfund  
Research  
Program

**Contact:** andres.cardenas@berkeley.edu; anne.bozack@berkeley.edu

**Funding:** NIEHS P42 ES004705, P42 ES010349



**NIEHSSRP**

**Berkeley**  
UNIVERSITY OF CALIFORNIA