# Article review: MCPS and *snipar*

B.Sc. Roberto Olvera-Hernandez

*Centre for Genomic Sciences (CCG),*
*National Autonomous University of Mexico (UNAM)*

**February 18, 2025**

CCG
Centro de Ciencias Genómicas

# Contents

# 1 Mexico City Prospective Study (MCPS)

# 1 Mexico City Prospective Study (MCPS)

## 1.1 Recruitment and baseline data

# Overview

Over **150,000 participants** were recruited in two districts between **1998 and 2004**.

- Baseline questionnaire.
- Blood samples.
- Physical measurements.
- Linkage to mortality.

COHORT PROFILE

**Cohort Profile: The Mexico City Prospective Study**

Roberto Tapia-Conyer,[1] Pablo Kuri-Morales,[2] Jesús Alegre-Díaz,[2] Gary Whitlock,[3]* Jonathan Emberson,[3] Sarah Clark,[3] Richard Peto[3] and Rory Collins[3]

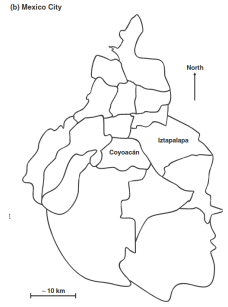


Figure 1: Map showing the location of the MCPS districts (Tapia-Conyer et al. 2006, *International Journal of Epidemiology*).

# Baseline data

## Socio-demographic

- Age and sex
- Area of residence
- Marital status
- Educational achievement
- Occupation
- Income
- Health service provider

# Baseline data

## Socio-demographic

- Age and sex
- Area of residence
- Marital status
- Educational achievement
- Occupation
- Income
- Health service provider

## Lifestyle characteristics

- Diet (fruit/vegetables, fried food, types of oil)
- Smoking and alcohol
- Physical activity
- Sleep duration

# Baseline data

## Socio-demographic

- Age and sex
- Area of residence
- Marital status
- Educational achievement
- Occupation
- Income
- Health service provider

## Lifestyle characteristics

- Diet (fruit/vegetables, fried food, types of oil)
- Smoking and alcohol
- Physical activity
- Sleep duration

## Reproductive history (women)

- Menopausal status
- Hysteroctomy
- Oopheroctomy
- HRT
- Contraceptive use
- Pregnancy (age and number)

# Baseline data

## Socio-demographic

- Age and sex
- Area of residence
- Marital status
- Educational achievement
- Occupation
- Income
- Health service provider

## Lifestyle characteristics

- Diet (fruit/vegetables, fried food, types of oil)
- Smoking and alcohol
- Physical activity
- Sleep duration

## Reproductive history (women)

- Menopausal status
- Hysteroctomy
- Oopheroctomy
- HRT
- Contraceptive use
- Pregnancy (age and number)

## Physical measurements

- Height
- Weight
- Waist and hip circumferece
- Systolic and diastolic blood pressure

# Baseline data

## Socio-demographic

- Age and sex
- Area of residence
- Marital status
- Educational achievement
- Occupation
- Income
- Health service provider

## Lifestyle characteristics

- Diet (fruit/vegetables, fried food, types of oil)
- Smoking and alcohol
- Physical activity
- Sleep duration

## Reproductive history (women)

- Menopausal status
- Hysteroctomy
- Oopheroctomy
- HRT
- Contraceptive use
- Pregnancy (age and number)

## Physical measurements

- Height
- Weight
- Waist and hip circumferece
- Systolic and diastolic blood pressure

## Blood samples

- Plasma & buffy coat
- HbA1c and other essays
- NMR metabolomics (e.g. fatty acids, cholines, lipoprotein subclasses, etc.)

# Baseline data

## Socio-demographic

- Age and sex
- Area of residence
- Marital status
- Educational achievement
- Occupation
- Income
- Health service provider

## Lifestyle characteristics

- Diet (fruit/vegetables, fried food, types of oil)
- Smoking and alcohol
- Physical activity
- Sleep duration

## Reproductive history (women)

- Menopausal status
- Hysteroctomy
- Oopheroctomy
- HRT
- Contraceptive use
- Pregnancy (age and number)

## Physical measurements

- Height
- Weight
- Waist and hip circumferece
- Systolic and diastolic blood pressure

## Blood samples

- Plasma & buffy coat
- HbA1c and other essays
- NMR metabolomics (e.g. fatty acids, cholines, lipoprotein subclasses, etc.)

## Prior diseases and medications

Participants were asked if they had ever been diagnosed with any of the listed diseases (binary: *Yes* or *No*).

# 1 Mexico City Prospective Study (MCPS)

## 1.2 Genetic overview

# Genetic datasets

- Genetic datasets were added later by Ziyatdinov et al. (2023), making it one of the **largest** studies for **non-eurpean** populations.

# Genetic datasets

- Genetic datasets were added later by Ziyatdinov et al. (2023), making it one of the **largest** studies for **non-eurpean** populations.
- Comparison (WES and WGS) were made with other datasets: **UK Biobank**, **TOPMed**, **gnomAD**.

# Genetic datasets

- Genetic datasets were added later by Ziyatdinov et al. (2023), making it one of the **largest** studies for **non-eurpean** populations.
- Comparison (WES and WGS) were made with other datasets: **UK Biobank**, **TOPMed**, **gnomAD**.

**Genome-Wide Genotyping**

- Illumina — GSAv2 chip array
- 138,511 individuals

# Genetic datasets

- Genetic datasets were added later by Ziyatdinov et al. (2023), making it one of the **largest** studies for **non-eurpean** populations.
- Comparison (WES and WGS) were made with other datasets: **UK Biobank**, **TOPMed**, **gnomAD**.

## Genome-Wide Genotyping

- Illumina — GSAv2 chip array
- 138,511 individuals

## Exome Sequencing (WES)

- $n = 141{,}046$ individuals

**Variants:**

- *Total:* 9.3 million.
- *Coding regions:* 4.0 million in 19,110 genes.
- *Unique MCPS:* 1.4 million.

# Genetic datasets

- Genetic datasets were added later by Ziyatdinov et al. (2023), making it one of the **largest** studies for **non-eurpean** populations.
- Comparison (WES and WGS) were made with other datasets: **UK Biobank**, **TOPMed**, **gnomAD**.

## Genome-Wide Genotyping

- Illumina — GSAv2 chip array
- 138,511 individuals

## Exome Sequencing (WES)

- $n = 141{,}046$ individuals

**Variants:**

- *Total:* 9.3 million.
- *Coding regions:* 4.0 million in 19,110 genes.
- *Unique MCPS:* 1.4 million.

## Whole-Genome Sequencing (WGS)

- $n = 9{,}950$ individuals

**Variants:**

- *Total:* 131.9 million.
- *Coding regions:* 1.5 million.
- *Unique MCPS:* 31.5 million.

# Genetic datasets

- Genetic datasets were added later by Ziyatdinov et al. (2023), making it one of the **largest** studies for **non-eurpean** populations.
- Comparison (WES and WGS) were made with other datasets: **UK Biobank**, **TOPMed**, **gnomAD**.

### Genome-Wide Genotyping

- Illumina — GSAv2 chip array
- 138,511 individuals

### Exome Sequencing (WES)

- $n = 141{,}046$ individuals

**Variants:**

- *Total:* 9.3 million.
- *Coding regions:* 4.0 million in 19,110 genes.
- *Unique MCPS:* 1.4 million.

### Whole-Genome Sequencing (WGS)

- $n = 9{,}950$ individuals

**Variants:**

- *Total:* 131.9 million.
- *Coding regions:* 1.5 million.
- *Unique MCPS:* 31.5 million.

- Both **WES** and **WGS** share 93.2% of the variants, with an increment of 2.3% on **WGS** data.

# WES and WGS - Comparisons

- Lower proportion of **singletons**, indicates *extensive familial relatedness*.
- Increased number of **predicted loss of function (pLOF)** variants.

| Variant Type | MAF | MCPS Freeze 150 WES All ancestries | | UKB WES All ancestries (N=454,787) | TOPMed Freeze 8[a] All ancestries (N=132,345) | gnomAD 3.1[a] All ancestries (N=76,156) |
|---|---|---|---|---|---|---|
| | | Total Variants | Unique to MCPS | | | |
| All coding[b] | All | 3,993,480 | 1,378,929 | 12,251,048 | 7,967,776 | 6,720,277 |
| | Singleton | 1,232,799 | 641,245 | 5,745,376 | 3,629,356 | 3,487,928 |
| | Doubleton - 0.01% | 2,249,180 | 718,941 | 6,105,074 | 3,755,427 | 2,619,225 |
| | 0.01-0.1% | 378,837 | 18,651 | 300,716 | 384,065 | 397,323 |
| | 0.1-1% | 81,522 | 46 | 54,860 | 118,746 | 129,469 |
| | 1-5% | 17,083 | 8 | 17,318 | 39,559 | 41,791 |
| | >5% | 34,059 | 38 | 27,704 | 40,623 | 44,541 |

# WES and WGS - Comparisons

- Lower proportion of **singletons**, indicates *extensive familial relatedness*.
- Increased number of **predicted loss of function (pLOF)** variants.

| Variant Type | MAF | MCPS WGS All ancestries (N=9950) | | TOPMed Freeze 8[a] | gnomAD 3.1[a] |
|---|---|---|---|---|---|
| | | Total Variants | Unique to MCPS | All ancestries (N=132,345) | All ancestries (N=76,156) |
| All Variants | All | 131,851,585 | 31,533,601 | 705,482,499 | 643,434,862 |
| | Singleton | 57,885,022 | 23,866,451 | 323,651,578 | 317,422,028 |
| | Doubleton-0.1% | 54,213,931 | 7,621,741 | 354,185,173 | 287,971,390 |
| | 0.1-1% | 10,811,295 | 33,835 | 14,418,471 | 20,236,181 |
| | 1-5% | 2,635,764 | 4,868 | 5,928,971 | 8,153,113 |
| | >5% | 6,305,573 | 6,706 | 7,298,306 | 9,652,150 |

# Family networks

**Estimation**

Relatedness was estimated through *identity-by-descent (IBD)* sharing.

# Family networks

**Estimation**

Relatedness was estimated through *identity-by-descent (IBD)* sharing.

About 71% of individuals have **at least one relative** present in the MCPS dataset.

- **Parent-Offspring (PO):** 31,597 relationships.
- **Sibling Pairs (FS):** 29,482 relationships.
- **Second Degree (2nd):** 47,080 relationships.
- **Third Degree (3rd):** 120,180 relationships.

# Family networks

**Estimation**

Relatedness was estimated through *identity-by-descent (IBD)* sharing.

About 71% of individuals have **at least one relative** present in the MCPS dataset.

- **Parent-Offspring (PO):** 31,597 relationships.
- **Sibling Pairs (FS):** 29,482 relationships.
- **Second Degree (2nd):** 47,080 relationships.
- **Third Degree (3rd):** 120,180 relationships.

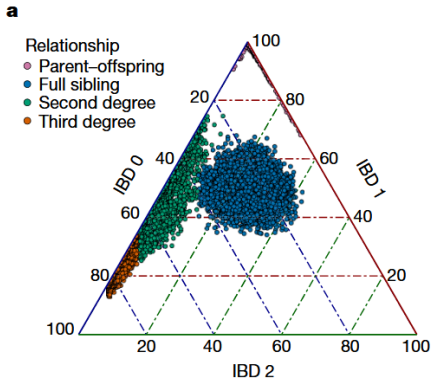


Figure 2: Percentage of genome estimated to have zero, one or two IBD alleles (Ziyatdinov et al. 2023, *Nature*).

# Family networks

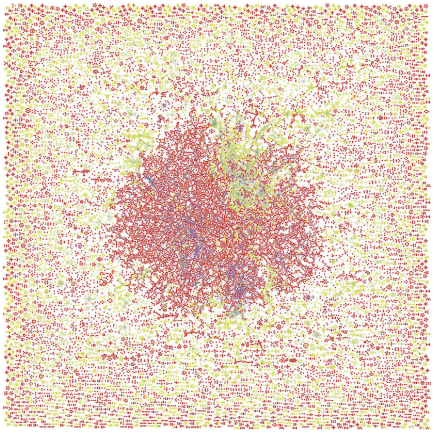

Figure 3: Graph of second-degree family networks of size four or greater (Ziyatdinov et al. 2023, *Nature*).

# Family networks

The levels of *relatedness* were:

- much higher than those from the **UK Biobank (UKB)**.
- comparable with the **Geisinger Health Study (GHS)**–both **MCPS** and **GHS** recruited in *close proximity*.

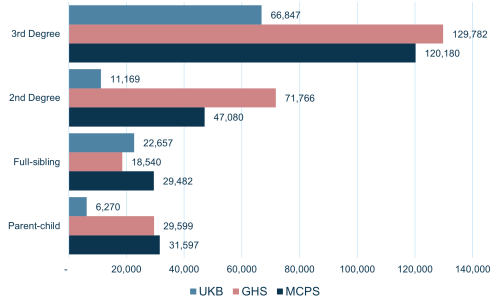


Figure 4: Comparison of network sizes in MCPS, UKB and GHS. Data extracted from Supplementary Table 25 (Ziyatdinov et al. 2023, *Nature*).

# 1 Mexico City Prospective Study (MCPS)

## 1.3 Population structure and ancestry

# Principal Components Analyses (PCA)

Characterization of ancestry composition adding *unrelated* samples to PCA

**1,000 Genomes Project**

- African (Yoruba): 108 samples
- European (Iberian): 107 samples

**Metabolic Analysis of an Indigenous Samples (MAIS)**

- Indigenous Mexican: 591 samples (60 populations)

(García-Ortiz et al. 2021, *Nature Communications*)

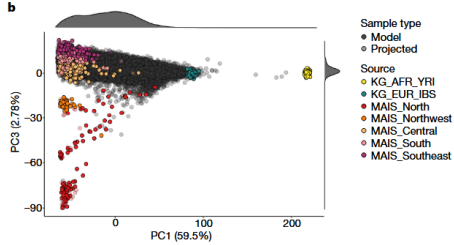


Figure 5: PCA for MCPS, African, European and Indigenous Mexican samples.

# References

García-Ortiz, H., Barajas-Olmos, F., Contreras-Cubas, C., Cid-Soto, M. Á., Córdova, E. J., Centeno-Cruz, F., Mendoza-Caamal, E., Cicerón-Arellano, I., Flores-Huacuja, M., Baca, P., Bolnick, D. A., Snow, M., Flores-Martínez, S. E., Ortiz-Lopez, R., Reynolds, A. W., Blanchet, A., Morales-Marín, M., Velázquez-Cruz, R., Kostic, A. D., … Orozco, L. (2021). *Nature Communications*, *12*(1), 5942. https://doi.org/10.1038/s41467-021-26188-w

Tapia-Conyer, R., Kuri-Morales, P., Alegre-Díaz, J., Whitlock, G., Emberson, J., Clark, S., Peto, R., & Collins, R. (2006). *International Journal of Epidemiology*, *35*(2), 243–249. https://doi.org/10.1093/ije/dyl042

Ziyatdinov, A., Torres, J., Alegre-Díaz, J., Backman, J., Mbatchou, J., Turner, M., Gaynor, S. M., Joseph, T., Zou, Y., Liu, D., Wade, R., Staples, J., Panea, R., Popov, A., Bai, X., Balasubramanian, S., Habegger, L., Lanche, R., Lopez, A., … Tapia-Conyer, R. (2023). *Nature*, *622*(7984), 784–793. https://doi.org/10.1038/s41586-023-06595-3