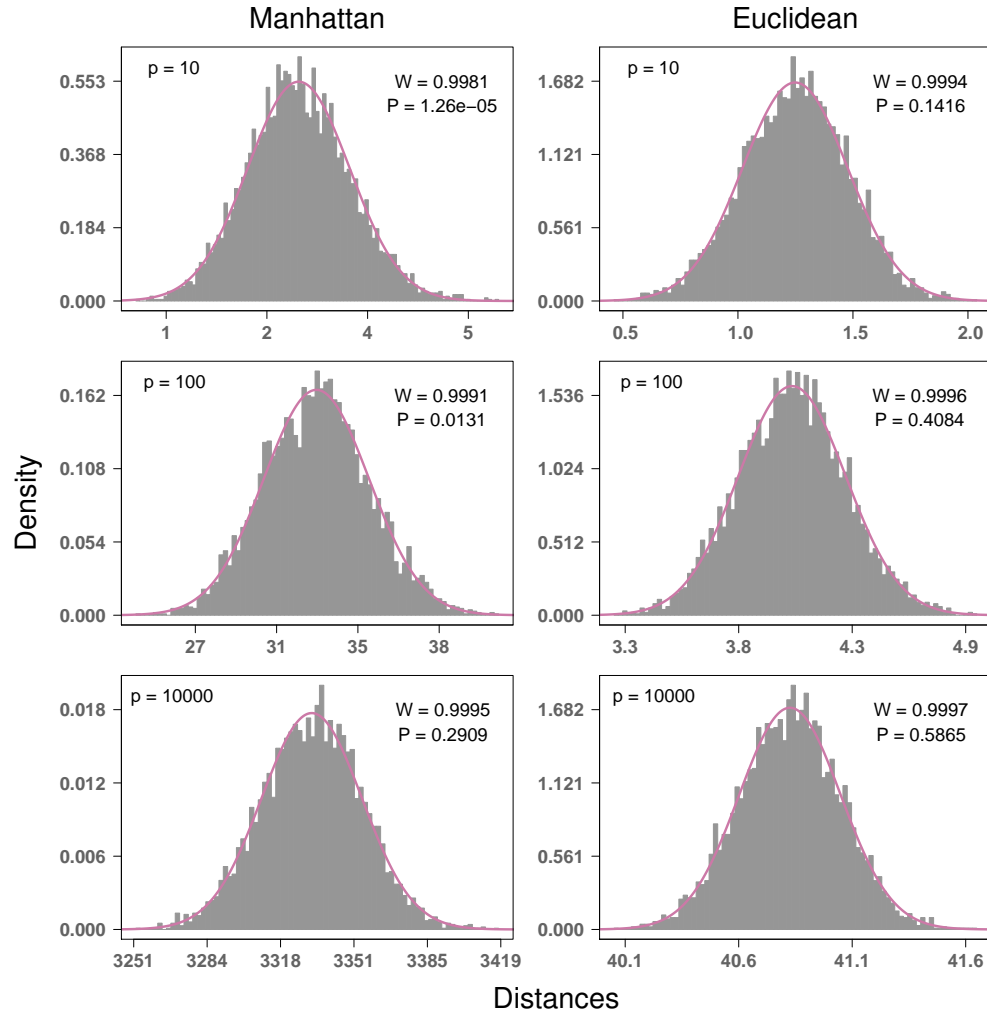
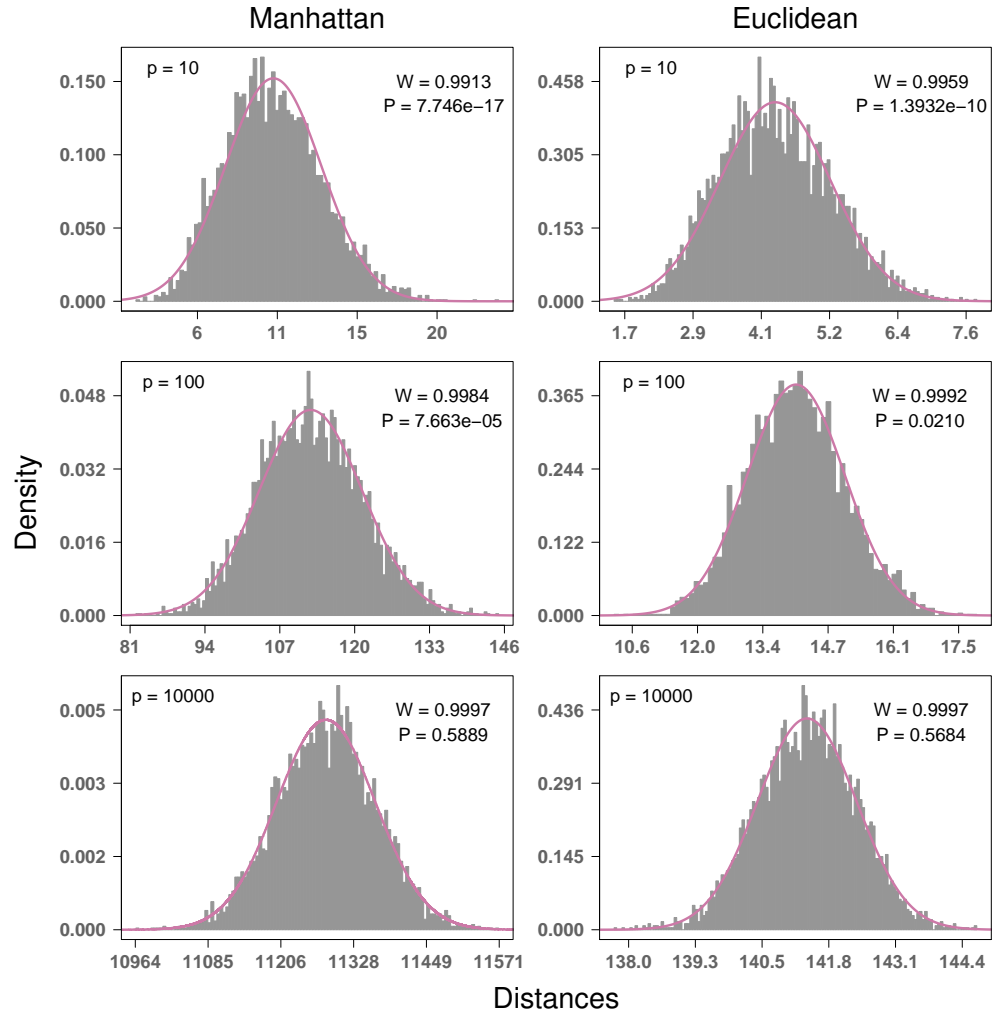


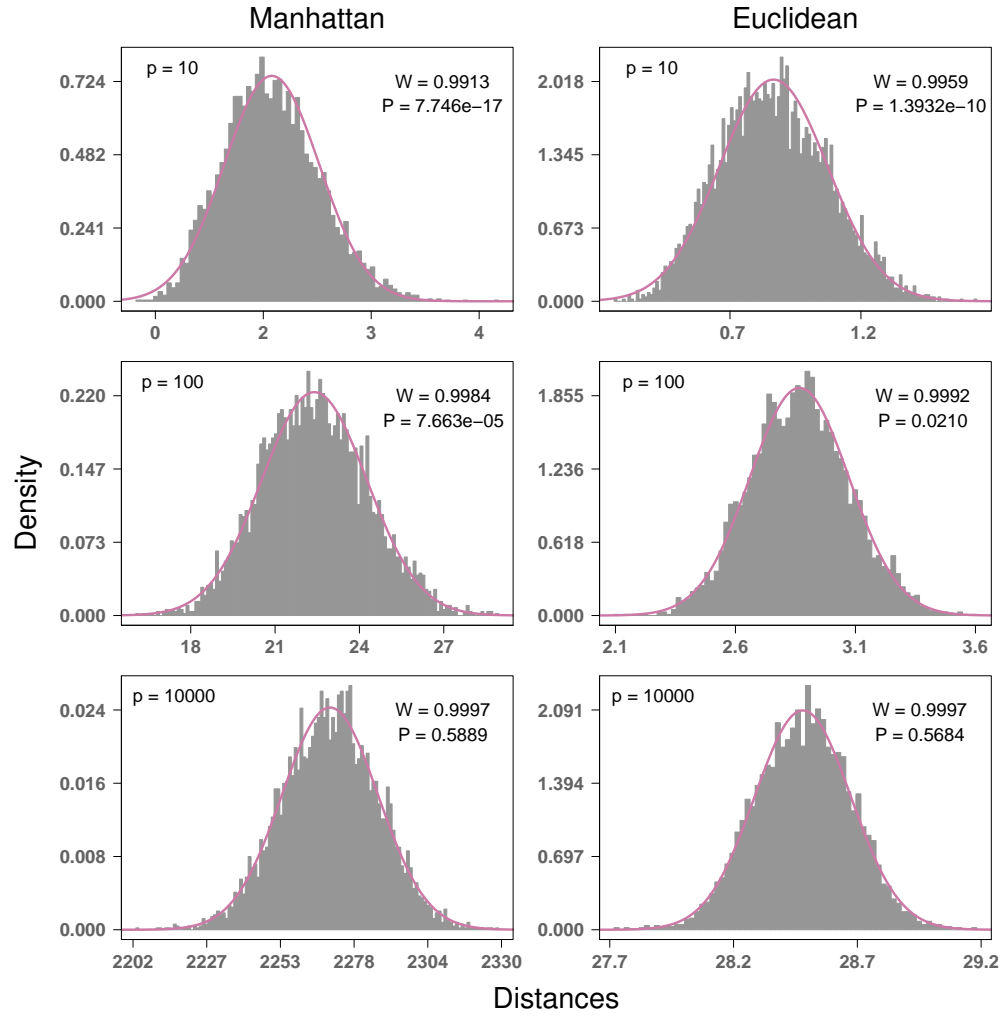
## Supplementary figures



**Figure S1.** Convergence to Gaussian for Manhattan and Euclidean distances for simulated standard uniform data with  $m = 100$  instances and  $p = 10, 100, \text{ and } 10000$  attributes. Convergence to Gaussian occurs rapidly with increasing  $p$ , and Gaussian is a good approximation for  $p$  as low as 10 attributes. The number of attributes in bioinformatics data is typically much larger, at least on the order of  $10^3$ . The Euclidean metric has stronger convergence to normal than Manhattan. P values from Shapiro-Wilk test, where the null hypothesis is a Gaussian distribution.

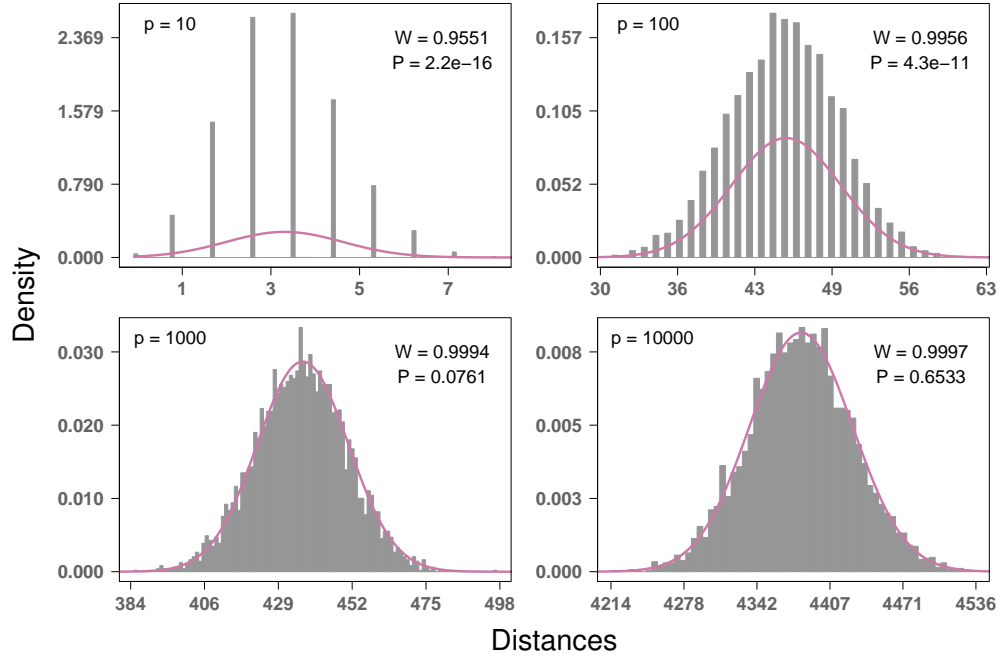


**Figure S2.** Convergence to Gaussian for Manhattan and Euclidean distances for simulated standard normal data with  $m = 100$  instances and  $p = 10, 100$ , and  $10000$  attributes. Convergence to Gaussian occurs rapidly with increasing  $p$ , and Gaussian is a good approximation for  $p$  as low as 10 attributes. The number of attributes in bioinformatics data is typically much larger, at least on the order of  $10^3$ . The Euclidean metric has stronger convergence to normal than Manhattan. P values from Shapiro-Wilk test, where the null hypothesis is a Gaussian distribution.



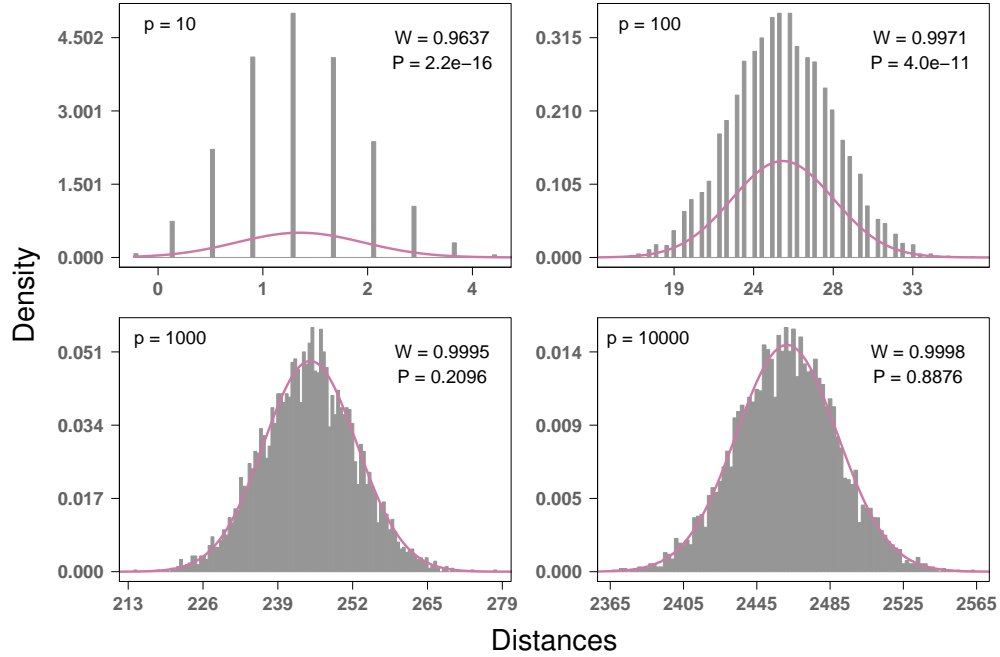
**Figure S3.** Convergence to Gaussian for max-min normalized Manhattan and Euclidean distances for simulated standard normal data with  $m = 100$  instances and  $p = 10, 100$ , and  $10000$  attributes. Convergence to Gaussian occurs rapidly with increasing  $p$ , and Gaussian is a good approximation for  $p$  as low as 10 attributes. The number of attributes in bioinformatics data is typically much larger, at least on the order of  $10^3$ . The Euclidean metric has stronger convergence to normal than Manhattan. P values from Shapiro-Wilk test, where the null hypothesis is a Gaussian distribution.

### Gaussian Convergence of GM Distances in GWAS Data



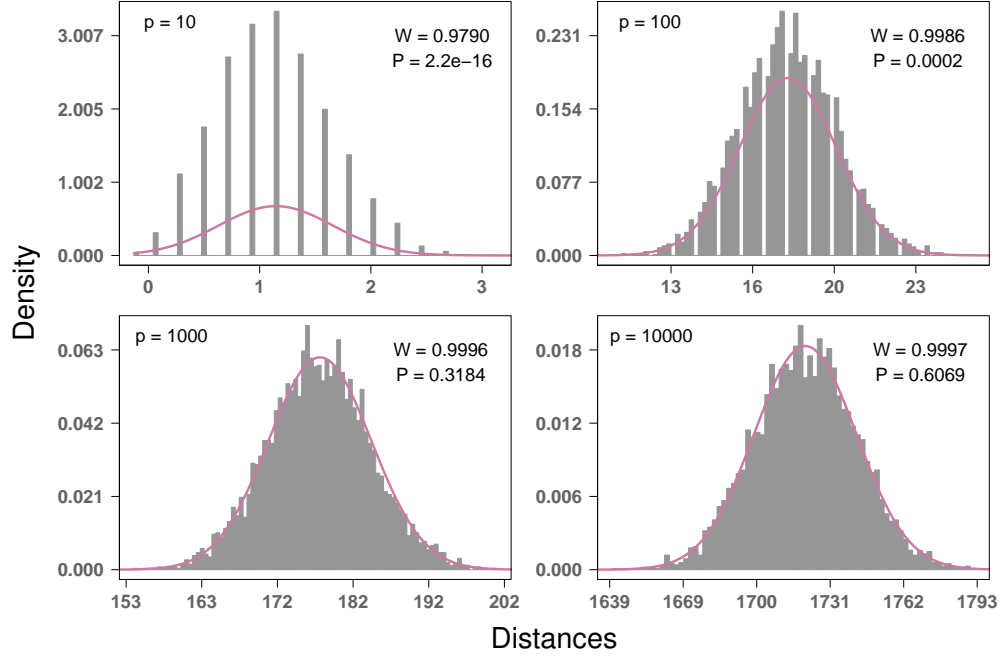
**Figure S4.** Convergence to Gaussian for GM distances for simulated binomial GWAS data with  $m = 100$  instances and  $p = 10, 100, 1000$ , and  $10000$  attributes. The average MAF was set to 0.205 for all simulations. Convergence to Gaussian occurs more gradually with increasing  $p$  than in continuous data. Significant convergence seems to occur when  $p \geq 1000$ , however, this is actually a relatively small number of features in the context of GWAS. Considering a realistic number of features for GWAS, the normality assumption of GM distances holds. This metric has the slowest convergence to Gaussian among all we have considered. P values from Shapiro-Wilk test, where the null hypothesis is a Gaussian distribution.

### Gaussian Convergence of AM Distances in GWAS Data



**Figure S5.** Convergence to Gaussian for AM distances for simulated binomial GWAS data with  $m = 100$  instances and  $p = 10, 100, 1000$ , and  $10000$  attributes. The average MAF was set to 0.205 for all simulations. Convergence to Gaussian occurs more gradually with increasing  $p$  than in continuous data. Significant convergence seems to occur when  $p \geq 1000$ , however, this is actually a relatively small number of features in the context of GWAS. Considering a realistic number of features for GWAS, the normality assumption of AM distances holds. This metric has the slightly faster convergence to Gaussian than the GM metric, which is probably due to the fact that the AM metric has one more value in its range (e.g.,  $1/2$ ). P values from Shapiro-Wilk test, where the null hypothesis is a Gaussian distribution.

### Gaussian Convergence of TiTv Distances in GWAS Data



**Figure S6.** Convergence to Gaussian for TiTv distances for simulated binomial GWAS data with  $m = 100$  instances and  $p = 10, 100, 1000$ , and  $10000$  attributes. The average MAF was set to 0.205 for all simulations and the Ti/Tv ratio ( $\eta$ ) was set to 2. Convergence to Gaussian occurs more gradually with increasing  $p$  than in continuous data. Significant convergence seems to occur when  $p \geq 1000$ , however, this is actually a relatively small number of features in the context of GWAS. Considering a realistic number of features for GWAS, the normality assumption of TiTv distances holds. This metric has the significantly faster convergence to Gaussian than the AM metric, which is probably due to the fact that the TiTv metric contains 2 more values in its range (e.g.,  $1/4$  &  $3/4$ ). P values from Shapiro-Wilk test, where the null hypothesis is a Gaussian distribution.

Figure 1 displays four histograms showing the distribution of distances for different sample sizes ( $p$ ), with a normal distribution fit curve overlaid on each. The x-axis represents Distances, and the y-axis represents Density.

- Top Left ( $p = 10$ ):** The distribution is centered around 105. The fit curve is a pink line. The statistical values are  $W = 0.9985$  and  $P = 0.0001$ .
- Top Right ( $p = 50$ ):** The distribution is centered around 2784. The fit curve is a pink line. The statistical values are  $W = 0.9996$  and  $P = 0.4214$ .
- Bottom Left ( $p = 150$ ):** The distribution is centered around 25304. The fit curve is a pink line. The statistical values are  $W = 0.9996$  and  $P = 0.3407$ .
- Bottom Right ( $p = 300$ ):** The distribution is centered around 101150. The fit curve is a pink line. The statistical values are  $W = 0.9997$  and  $P = 0.5533$ .

7/20

Figure 2 displays four histograms showing the distribution of distances for different values of  $p$  (10, 50, 150, and 300). Each plot includes a grey histogram, a pink normal distribution curve, and statistical values  $W$  and  $P$ .

- Top Left ( $p = 10$ ):** The x-axis ranges from 11 to 19, and the y-axis (Density) ranges from 0.000 to 0.287.  $W = 0.9985$ ,  $P = 0.0001$ .
- Top Right ( $p = 50$ ):** The x-axis ranges from 350 to 395, and the y-axis (Density) ranges from 0.000 to 0.061.  $W = 0.9996$ ,  $P = 0.5309$ .
- Bottom Left ( $p = 150$ ):** The x-axis ranges from 3113 to 3231, and the y-axis (Density) ranges from 0.000 to 0.024.  $W = 0.9996$ ,  $P = 0.4321$ .
- Bottom Right ( $p = 300$ ):** The x-axis ranges from 12067 to 12305, and the y-axis (Density) ranges from 0.000 to 0.012.  $W = 0.9997$ ,  $P = 0.4653$ .

The x-axis for all plots is labeled "Distances" and the y-axis is labeled "Density".

8/20



**A**

Theoretical Mean

Simulated Mean

$p = 1000$

$p = 2000$

$p = 3000$

$p = 4000$

$p = 5000$

**B**

Theoretical SD

Simulated SD

$p = 1000$

$p = 2000$

$p = 3000$

$p = 4000$

$p = 5000$

**Figure S9.** This will be a caption. This will be a caption. This will be a caption. This  
will be a caption. This will be a caption. This will be a caption. This will be a caption.  
This will be a caption. This will be a caption. This will be a caption. This will be a  
caption. This will be a caption. This will be a caption. This will be a caption. This will  
be a caption. This will be a caption. This will be a caption. This will be a caption.  
This will be a caption. This will be a caption. This will be a caption. This will be a  
caption. This will be a caption. This will be a caption. This will be a caption. This will  
be a caption. This will be a caption. This will be a caption. This will be a caption.  
This will be a caption. This will be a caption. This will be a caption. This will be a  
caption. This will be a caption. This will be a caption. This will be a caption. This will  
be a caption. This will be a caption.

**A**

Theoretical Mean

Simulated Mean

$p = 1000$

$p = 2000$

$p = 3000$

$p = 4000$

$p = 5000$

**B**

Theoretical SD

Simulated SD

$p = 1000$

$p = 2000$

$p = 3000$

$p = 4000$

$p = 5000$

10/20

**A**

Theoretical Mean

Simulated Mean

$p = 1000$

$p = 2000$

$p = 3000$

$p = 4000$

$p = 5000$

**B**

Theoretical SD

Simulated SD

$p = 1000$

$p = 2000$

$p = 3000$

$p = 4000$

$p = 5000$

October 15, 2019

Figure 1 consists of two panels, A and B, showing the relationship between theoretical and simulated values for different p values.

**Panel A: Theoretical Mean vs. Simulated Mean**

The x-axis is 'Simulated Mean' (ranging from 10 to 22) and the y-axis is 'Theoretical Mean' (ranging from 10 to 20). A dashed purple line represents the theoretical distribution. Data points are plotted for p = 1000 (purple), p = 2000 (blue), p = 3000 (teal), p = 4000 (green), and p = 5000 (light green). The points follow the dashed line, indicating that the simulated mean closely matches the theoretical mean.

p value	Simulated Mean	Theoretical Mean
1000	~9.5	~9.5
2000	~12.5	~12.5
3000	~15.5	~15.5
4000	~18.5	~18.5
5000	~21.5	~21.5

**Panel B: Theoretical SD vs. Simulated SD**

The x-axis is 'Simulated SD' (ranging from 0.198 to 0.200) and the y-axis is 'Theoretical SD' (ranging from 0.198 to 0.200). Data points are plotted for p = 1000 (purple), p = 2000 (blue), p = 3000 (teal), p = 4000 (green), and p = 5000 (light green). The points are clustered around the theoretical SD of 0.200, indicating that the simulated SD closely matches the theoretical SD.

p value	Simulated SD	Theoretical SD
1000	~0.1998	0.200
2000	~0.1999	0.200
3000	~0.2000	0.200
4000	~0.2001	0.200
5000	~0.2002	0.200

12/20

**A**

Theoretical Mean

Simulated Mean

$p = 1000$

$p = 2000$

$p = 3000$

$p = 4000$

$p = 5000$

**B**

Theoretical SD

Simulated SD

$p = 1000$

$p = 2000$

$p = 3000$

$p = 4000$

$p = 5000$

October 15, 2019

Figure 1 consists of two panels, A and B, showing the relationship between theoretical and simulated values for the proposed test across different values of  $p$  (1000, 2000, 3000, 4000, 5000).

**Panel A: Theoretical Mean vs. Simulated Mean**

The x-axis represents the Simulated Mean (ranging from 60 to 120), and the y-axis represents the Theoretical Mean (ranging from 40 to 100). A dashed purple line indicates the theoretical relationship. The data points are plotted for each  $p$  value, showing a positive correlation between the simulated and theoretical means.

$p$	Simulated Mean	Theoretical Mean
1000	~55	~45
2000	~65	~65
3000	~78	~78
4000	~90	~90
5000	~100	~100

**Panel B: Theoretical SD vs. Simulated SD**

The x-axis represents the Simulated SD (ranging from 0.995 to 1.005), and the y-axis represents the Theoretical SD (ranging from 0.990 to 1.000). The data points are plotted for each  $p$  value, showing a negative correlation between the simulated and theoretical standard deviations.

$p$	Simulated SD	Theoretical SD
1000	~1.000	~1.000
2000	~1.000	~0.992
3000	~1.000	~0.985
4000	~1.000	~0.996
5000	~1.000	~0.980

[illegible]

Figure 1 consists of two panels, A and B, showing the relationship between theoretical and simulated values for different p values.

**Panel A: Theoretical Mean vs. Simulated Mean**

The x-axis is 'Simulated Mean' (ranging from 10 to 22) and the y-axis is 'Theoretical Mean' (ranging from 10 to 20). A dashed purple line represents the theoretical distribution. Data points are plotted for p = 1000 (purple), p = 2000 (blue), p = 3000 (teal), p = 4000 (green), and p = 5000 (light green). The points follow the dashed line, indicating that the simulated mean closely matches the theoretical mean.

p value	Simulated Mean	Theoretical Mean
1000	~9.5	~9.5
2000	~12.5	~12.5
3000	~15.5	~15.5
4000	~18.5	~18.5
5000	~21.5	~21.5

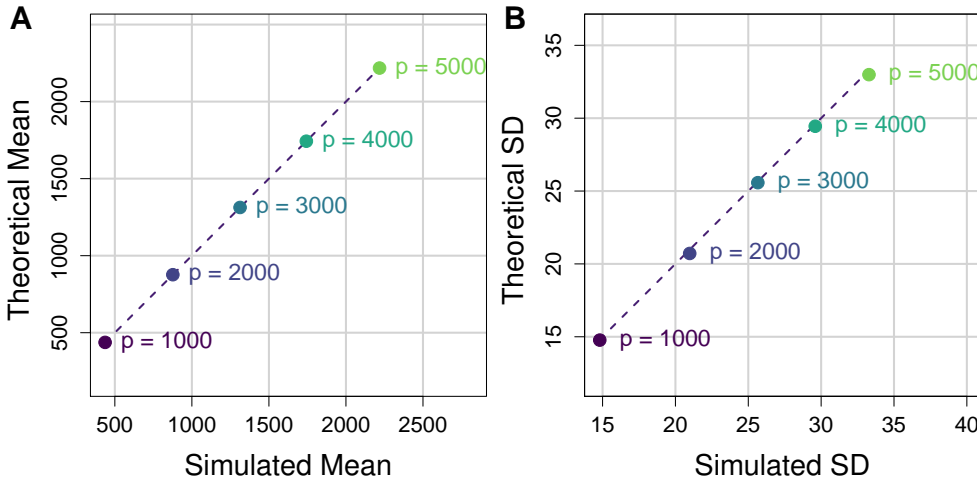
**Panel B: Theoretical SD vs. Simulated SD**

The x-axis is 'Simulated SD' (ranging from 0.198 to 0.200) and the y-axis is 'Theoretical SD' (ranging from 0.198 to 0.200). Data points are plotted for p = 1000 (purple), p = 2000 (blue), p = 3000 (teal), p = 4000 (green), and p = 5000 (light green). The points are clustered around the theoretical SD of 0.200, indicating that the simulated SD closely matches the theoretical SD.

p value	Simulated SD	Theoretical SD
1000	~0.1998	0.200
2000	~0.1999	0.200
3000	~0.2000	0.200
4000	~0.2001	0.200
5000	~0.2002	0.200

[illegible]

## Moments of GM Distances in GWAS Data



**Figure S16.** This will be a caption. This will be a caption. This will be a caption.  
This will be a caption. This will be a caption. This will be a caption. This will be a  
caption. This will be a caption. This will be a caption. This will be a caption. This will  
be a caption. This will be a caption. This will be a caption. This will be a caption.  
This will be a caption. This will be a caption. This will be a caption. This will be a  
caption. This will be a caption. This will be a caption. This will be a caption. This will  
be a caption. This will be a caption. This will be a caption. This will be a caption.  
This will be a caption. This will be a caption. This will be a caption. This will be a  
caption. This will be a caption. This will be a caption. This will be a caption. This will  
be a caption. This will be a caption. This will be a caption. This will be a caption. This will  
be a caption. This will be a caption. This will be a caption. This will be a caption.  
This will be a caption. This will be a caption.



**A**

Theoretical Mean

Simulated Mean

$p = 1000$

$p = 2000$

$p = 3000$

$p = 4000$

$p = 5000$

**B**

Theoretical SD

Simulated SD

$p = 1000$

$p = 2000$

$p = 3000$

$p = 4000$

$p = 5000$

October 15, 2019

**A**

Theoretical Mean

Simulated Mean

$p = 1000$

$p = 2000$

$p = 3000$

$p = 4000$

$p = 5000$

**B**

Theoretical SD

Simulated SD

$p = 1000$

$p = 2000$

$p = 3000$

$p = 4000$

$p = 5000$

October 15, 2019

**A**

Theoretical Mean

Simulated Mean

$p = 1000$

$p = 2000$

$p = 3000$

$p = 4000$

$p = 5000$

**B**

Theoretical SD

Simulated SD

$p = 1000$

$p = 2000$

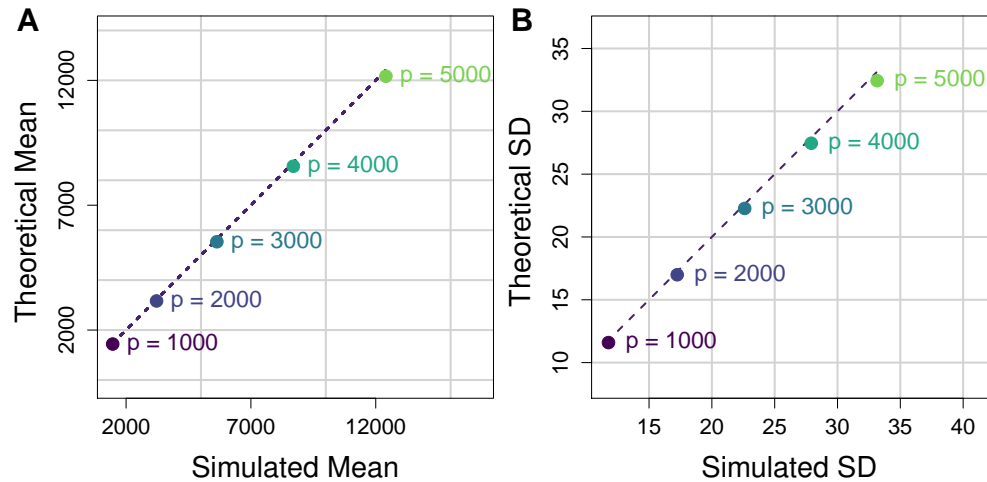
$p = 3000$

$p = 4000$

$p = 5000$

October 15, 2019

### Moments of max-min Normalized rs-fMRI Distances



**Figure S20.** This will be a caption. This will be a caption. This will be a caption.  
This will be a caption. This will be a caption. This will be a caption. This will be a  
caption. This will be a caption. This will be a caption. This will be a caption. This will  
be a caption. This will be a caption. This will be a caption. This will be a caption.  
This will be a caption. This will be a caption. This will be a caption. This will be a  
caption. This will be a caption. This will be a caption. This will be a caption. This will  
be a caption. This will be a caption. This will be a caption. This will be a caption.  
This will be a caption. This will be a caption. This will be a caption. This will be a  
caption. This will be a caption. This will be a caption. This will be a caption. This will  
be a caption. This will be a caption. This will be a caption. This will be a caption.  
This will be a caption. This will be a caption.