Supplementary figures

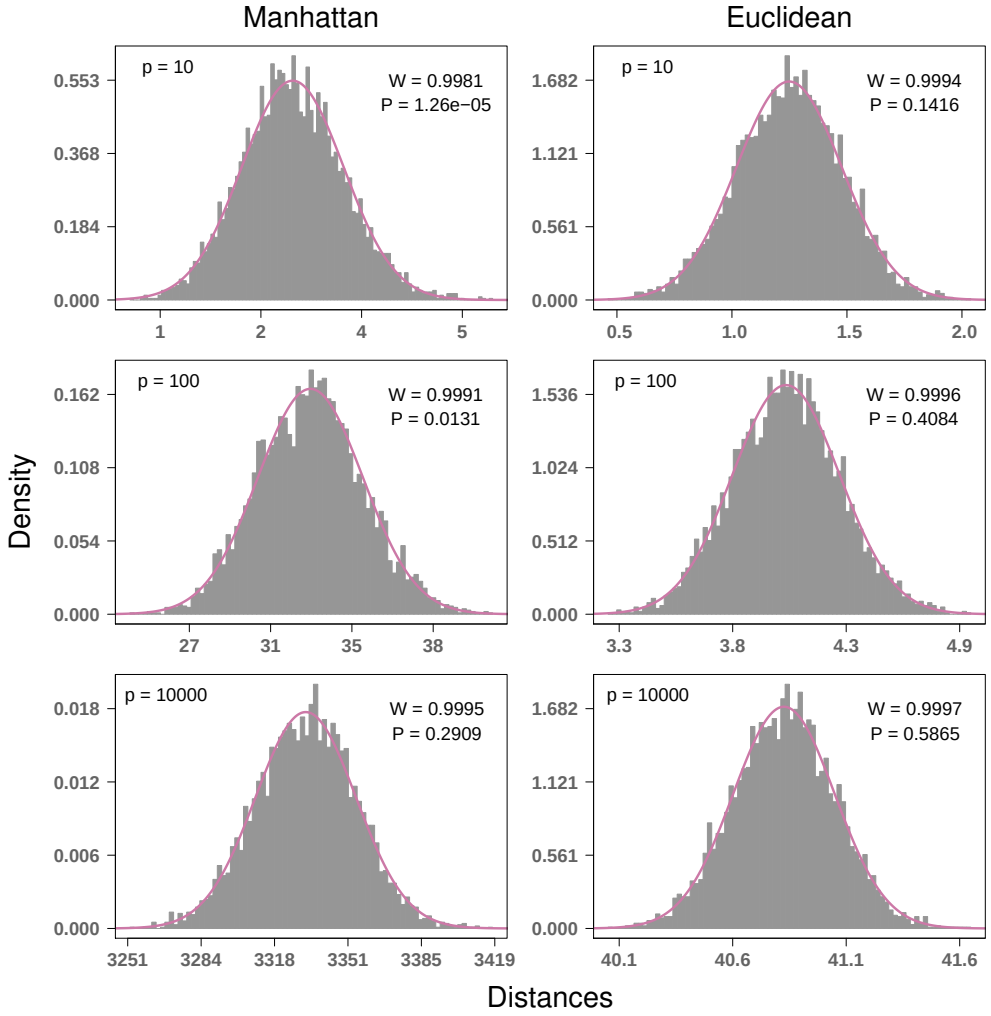

Figure S1. Convergence to Gaussian for Manhattan and Euclidean distances for simulated standard uniform data with $m = 100$ instances and $p = 10, 100$, and 10000 attributes. Convergence to Gaussian occurs rapidly with increasing p, and Gaussian is a good approximation for p as low as 10 attributes. 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.

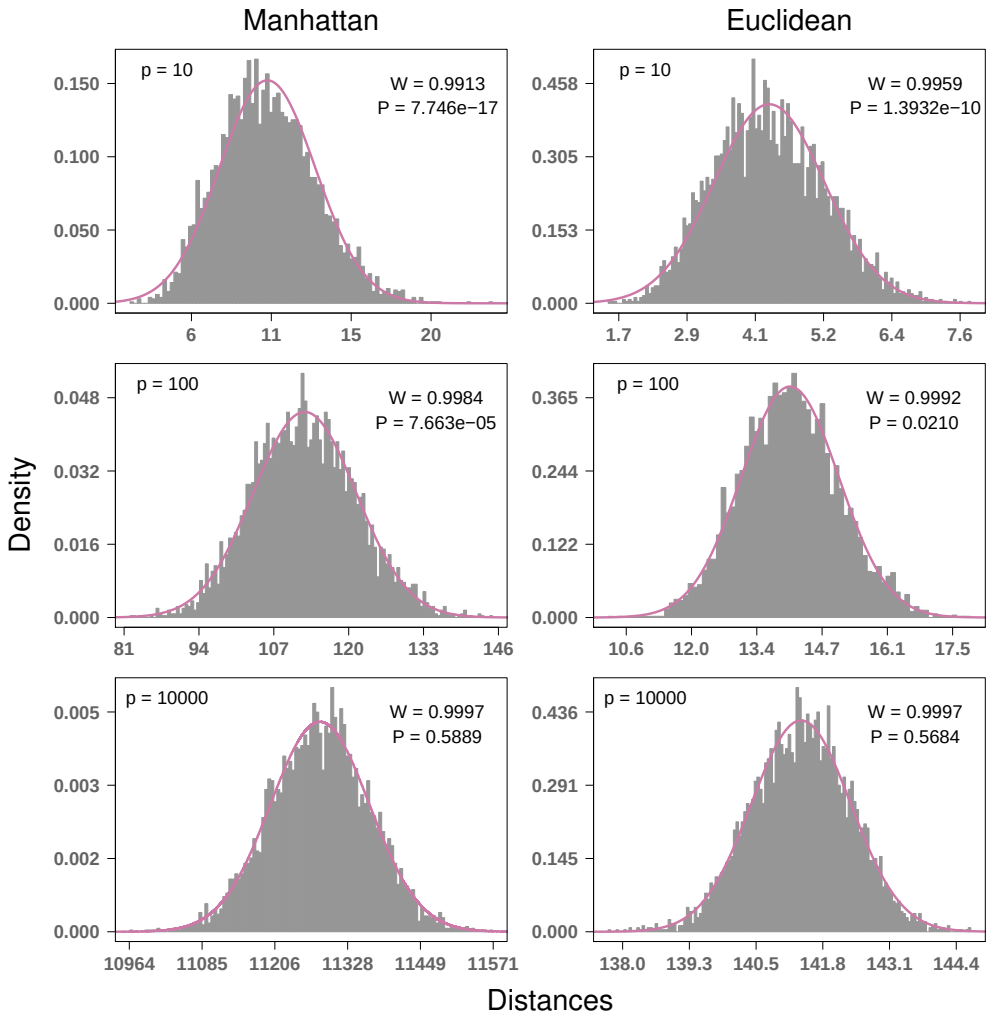

Figure S2. Convergence to Gaussian for Manhattan and Euclidean distances for simulated standard normal data with $m = 100$ instances and $p = 10, 100$, and 10000 attributes. Convergence to Gaussian occurs rapidly with increasing p, and Gaussian is a good approximation for p as low as 10 attributes. The 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.

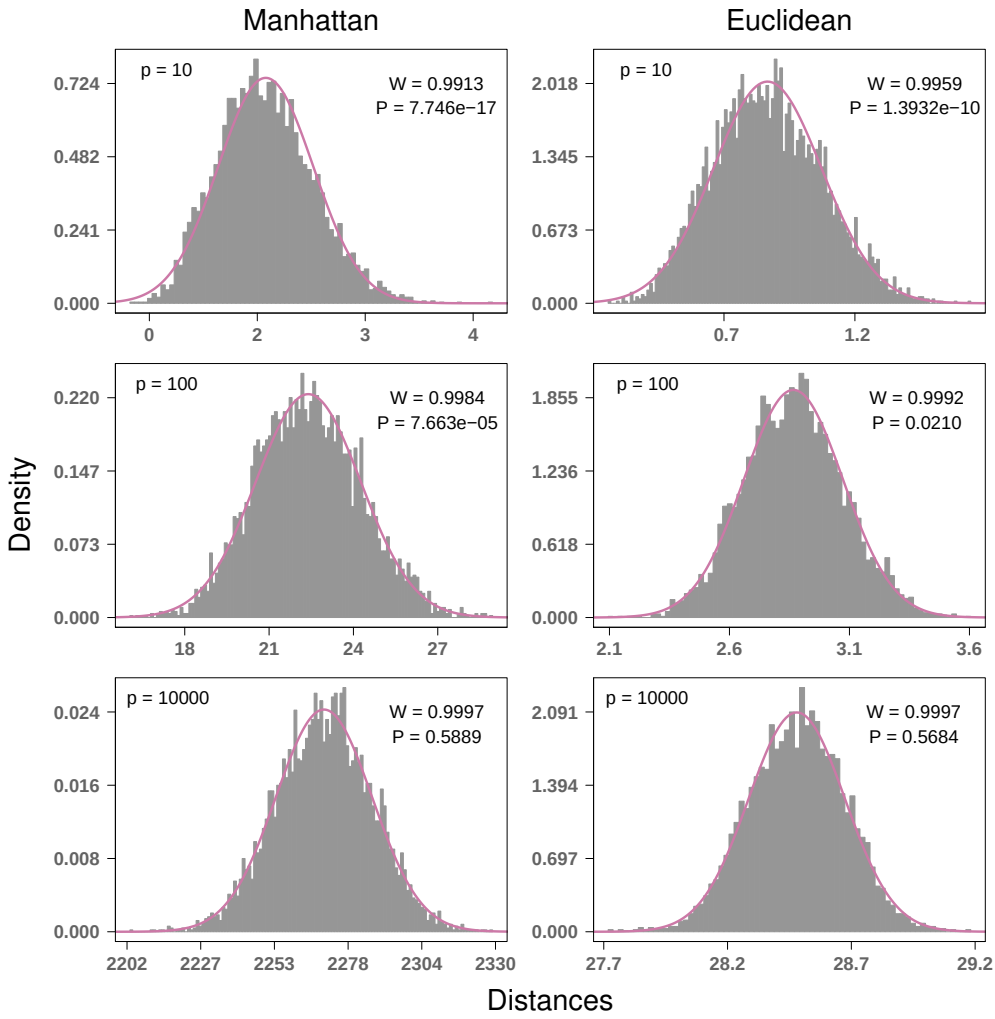

Figure S3. Convergence to Gaussian for max-min normalized Manhattan and Euclidean distances for simulated standard normal data with $m = 100$ instances and $p = 10, 100$, and 10000 attributes. Convergence to Gaussian occurs rapidly with increasing p, and Gaussian is a good approximation for p as low 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.

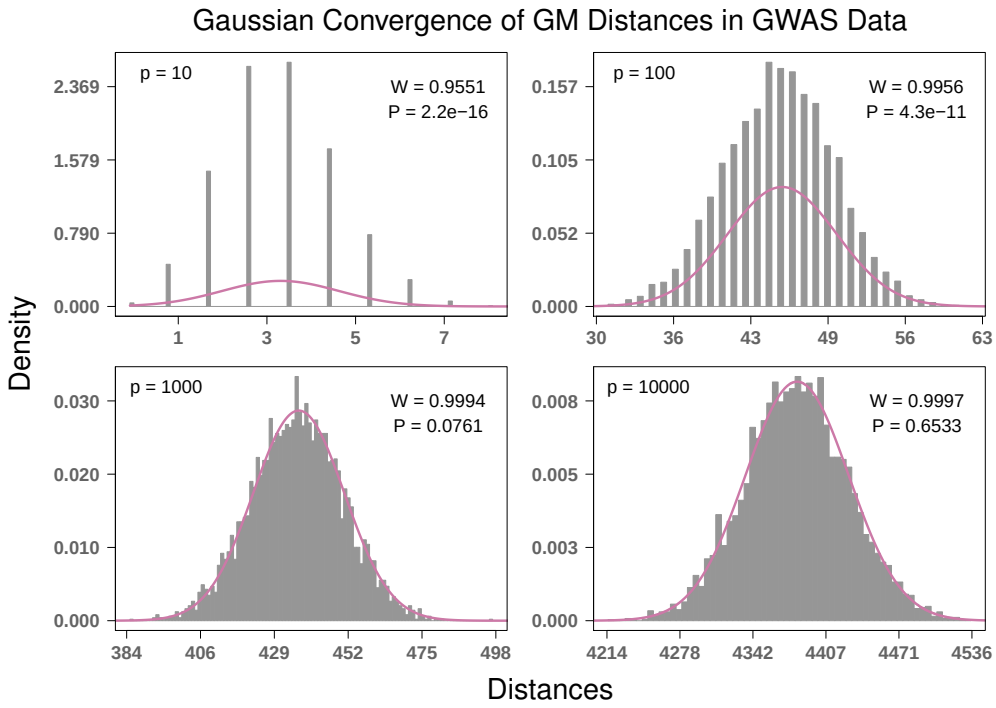

Figure S4. Convergence to Gaussian for GM distances for simulated binomial GWAS data with $m = 100$ instances and $p = 10, 100, 1000$, and 10000 attributes. The average MAF was set to 0.205 for all simulations. Convergence to Gaussian occurs more gradually with increasing p than in continuous data. Significant 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.

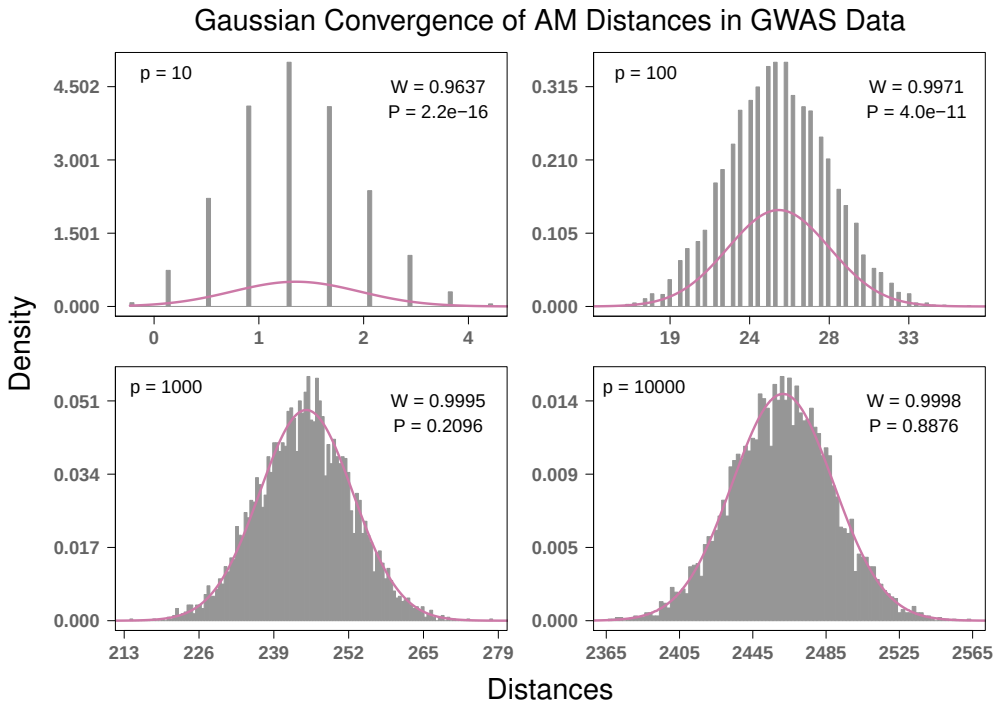

Figure S5. Convergence to Gaussian for AM distances for simulated binomial GWAS data with $m = 100$ instances and $p = 10, 100, 1000$, and 10000 attributes. The average MAF was set to 0.205 for all simulations. Convergence to Gaussian occurs more gradually with increasing p than in continuous data. Significant 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.

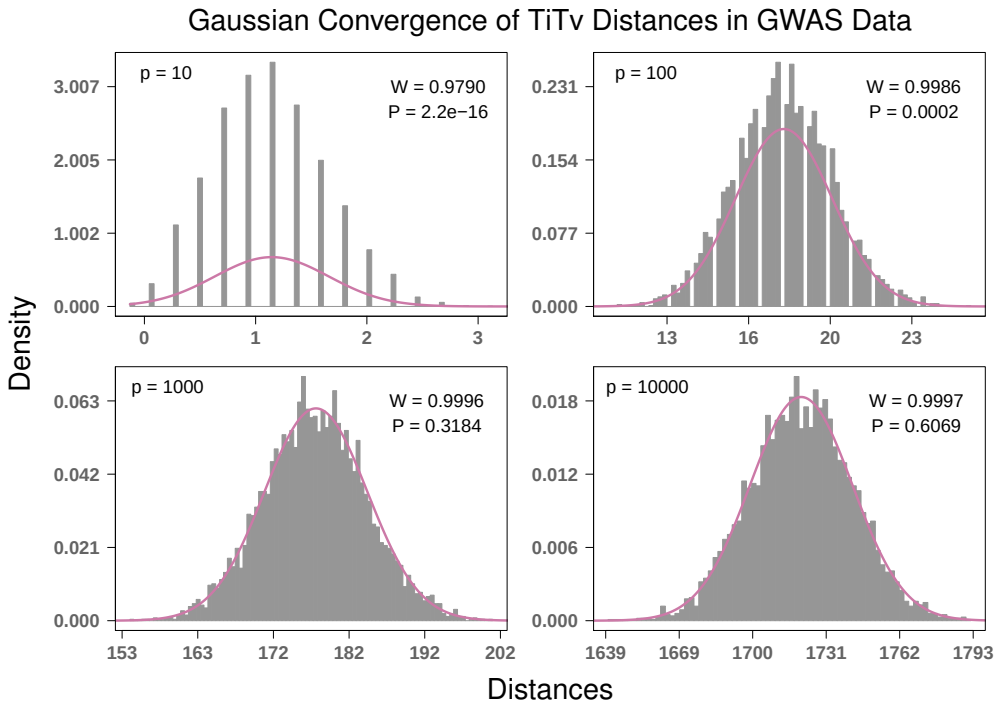

Figure S6. Convergence to Gaussian for TiTv distances for simulated binomial GWAS data with $m = 100$ instances and $p = 10, 100, 1000$, and 10000 attributes. The average MAF was set to 0.205 for all simulations and the Ti/Tv ratio (η) was set to 2. Convergence to Gaussian occurs more gradually with 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.

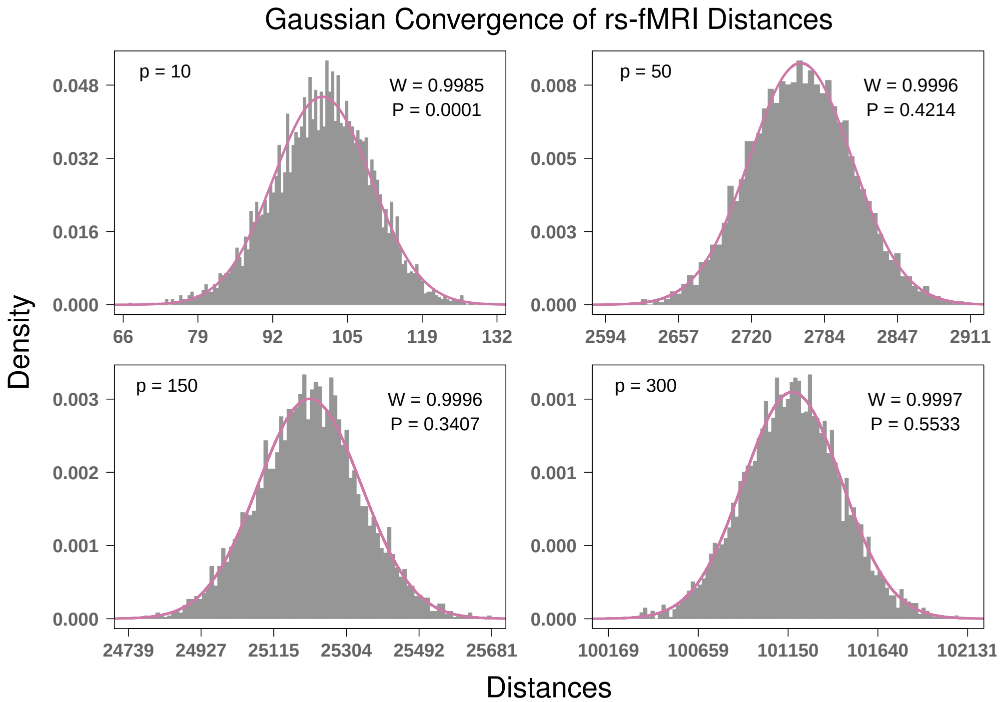

Figure S7. This will be a caption.

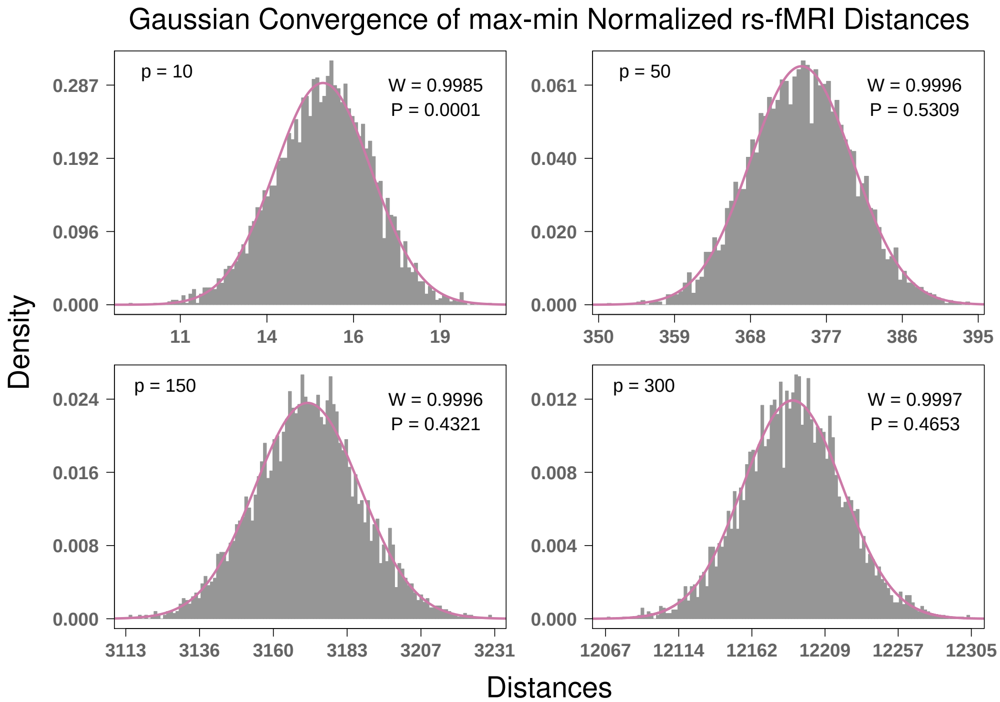

Figure S8. This will be a caption.

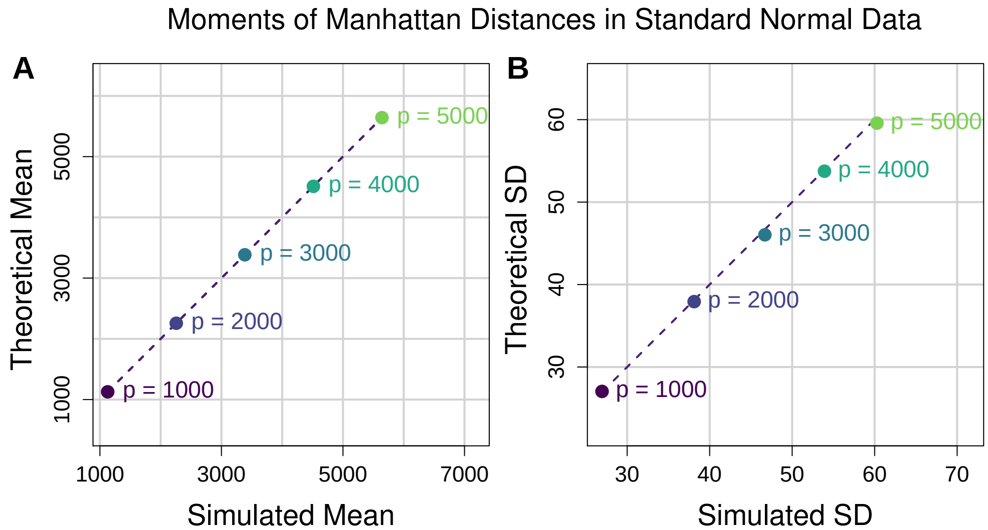

Figure S9. This will be a caption.

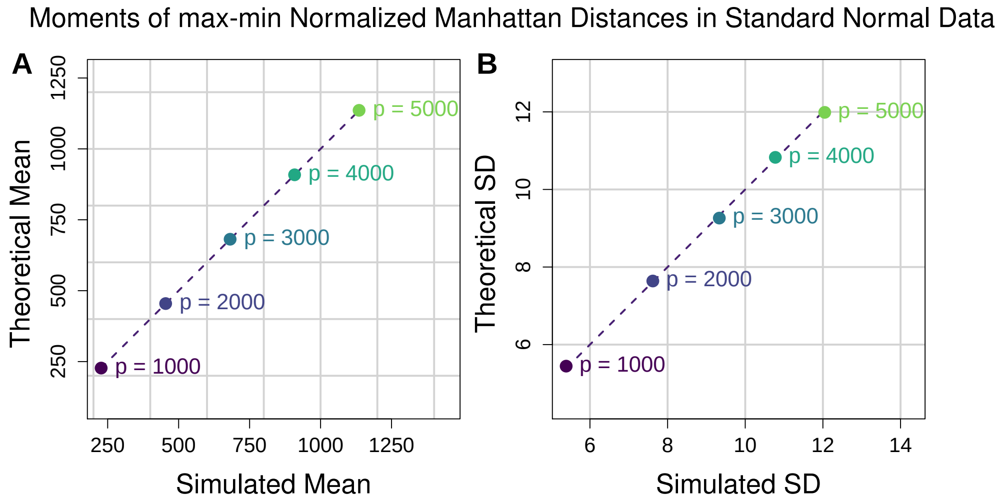

Figure S10. This will be a caption.

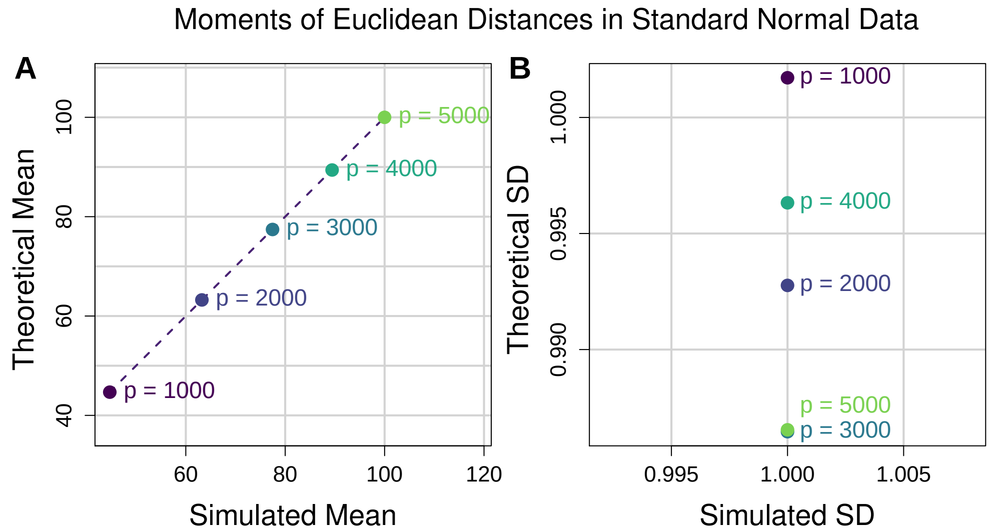

Figure S11. This will be a caption.

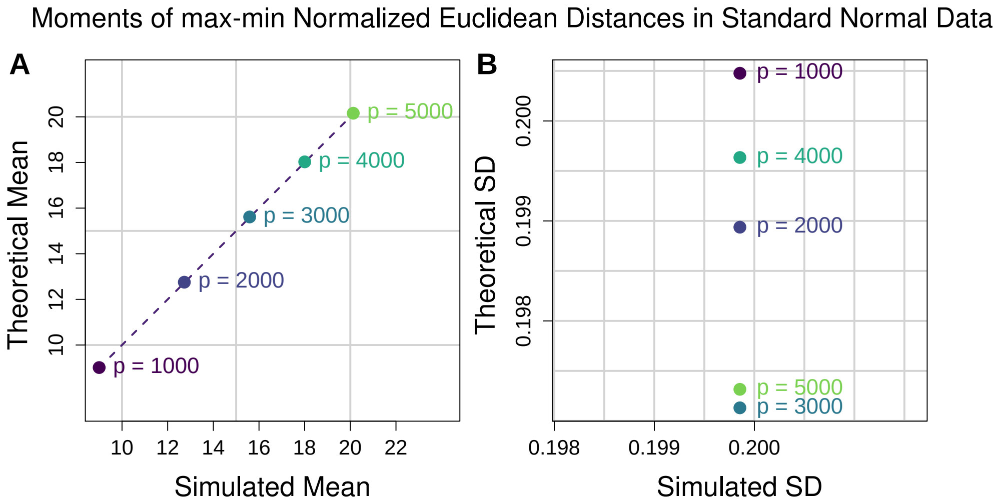

Figure S12. This will be a caption.

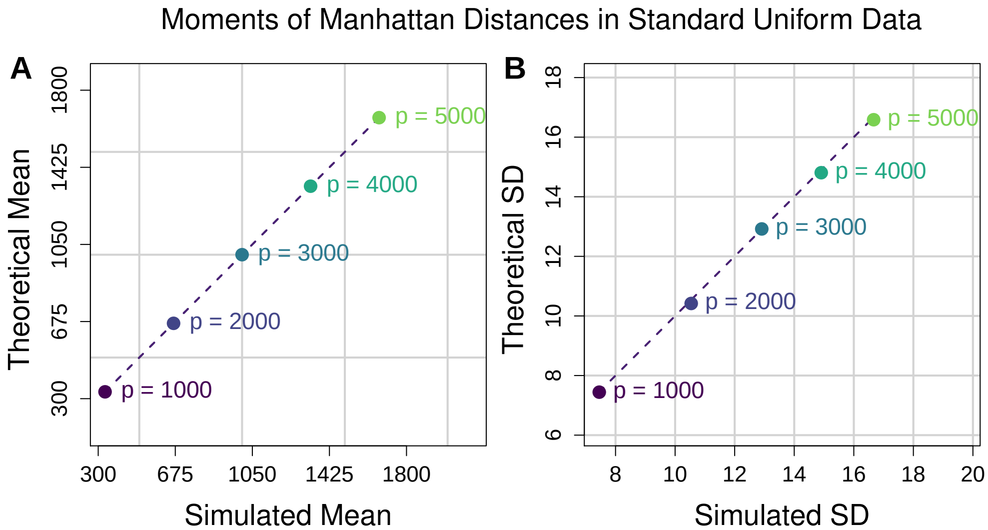

Figure S13. This will be a caption.

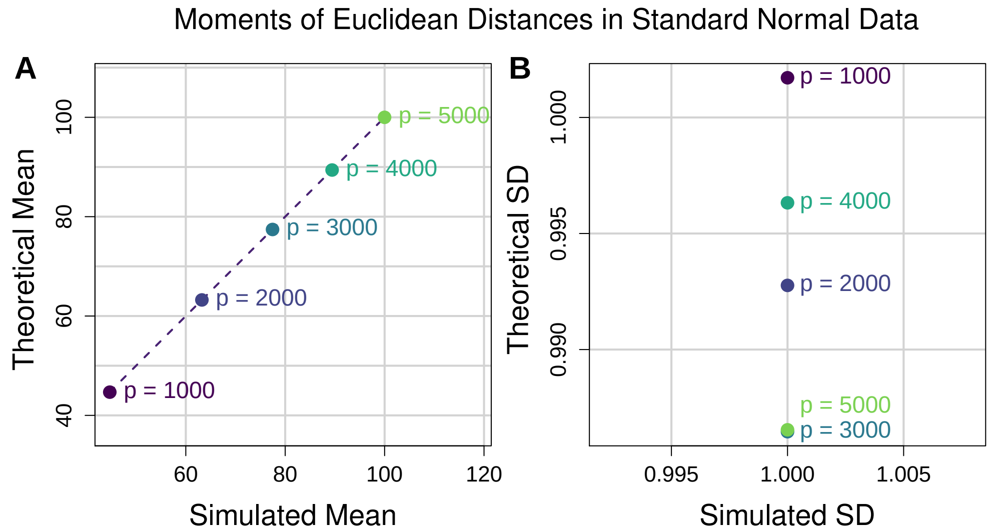

Figure S14. This will be a caption.

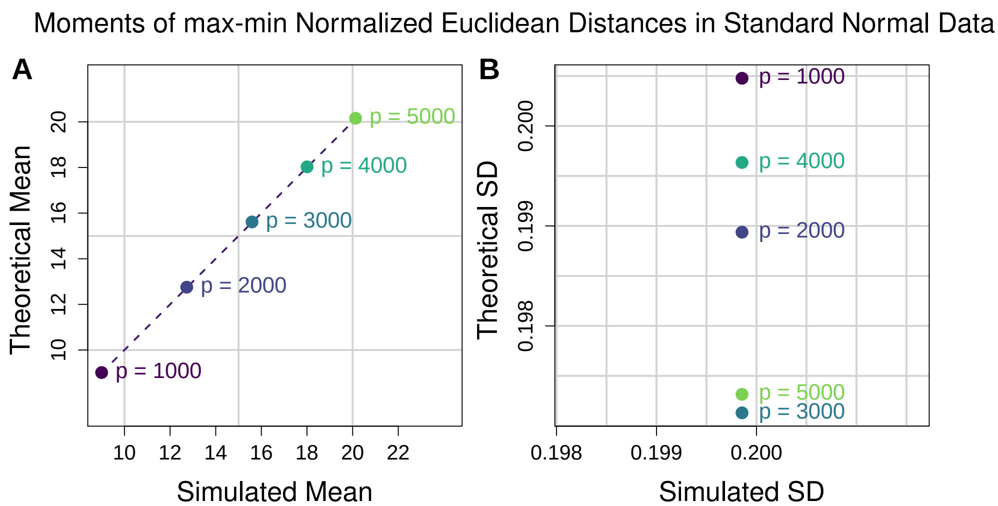

Figure S15. This will be a caption.

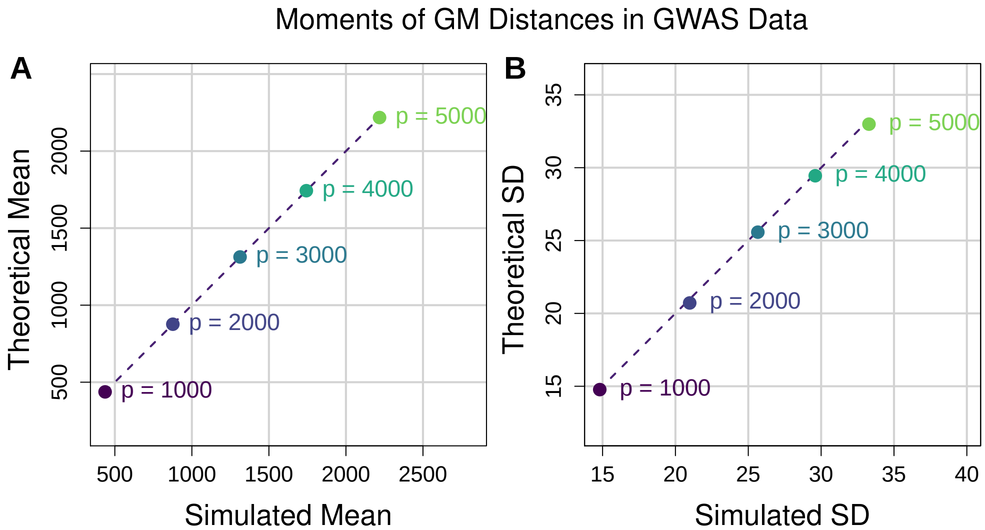

Figure S16. This will be a caption.

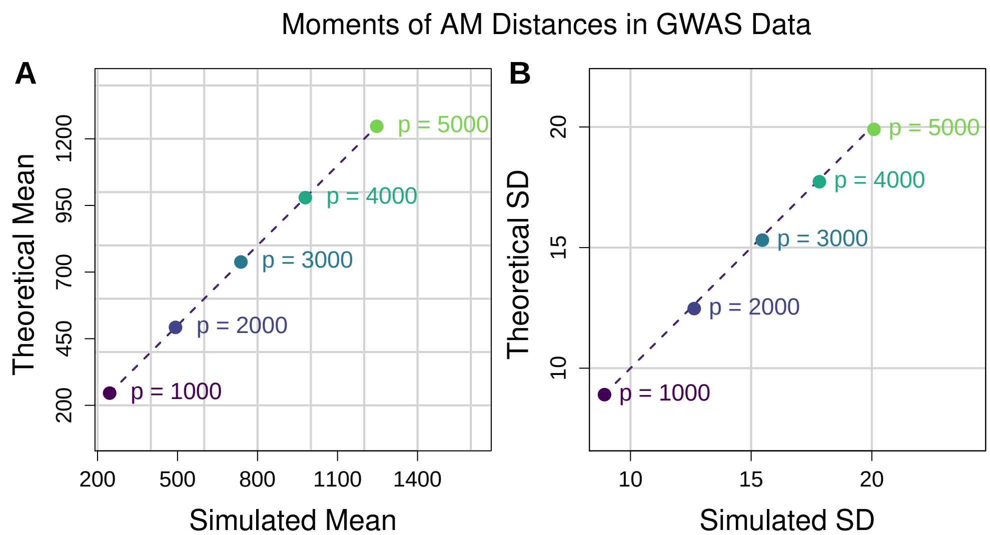

Figure S17. This will be a caption.

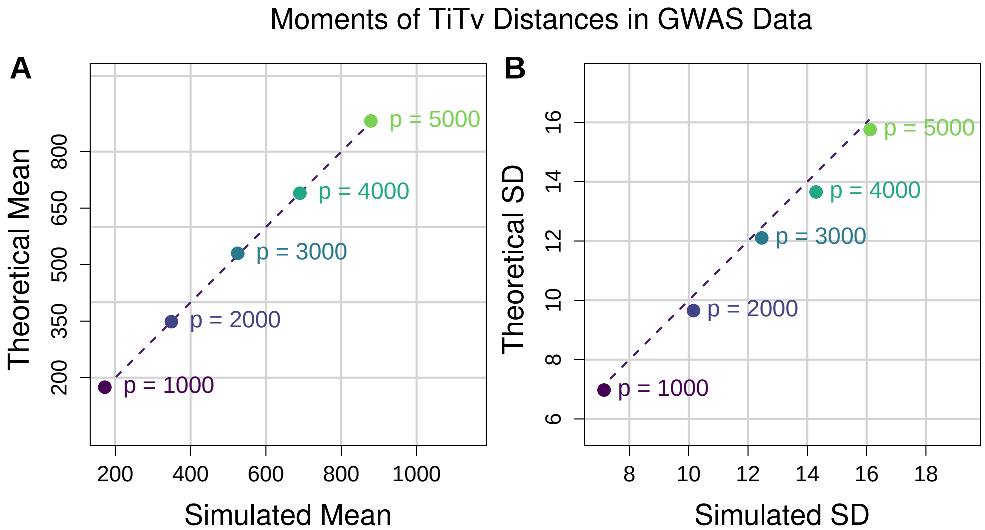

Figure S18. This will be a caption.

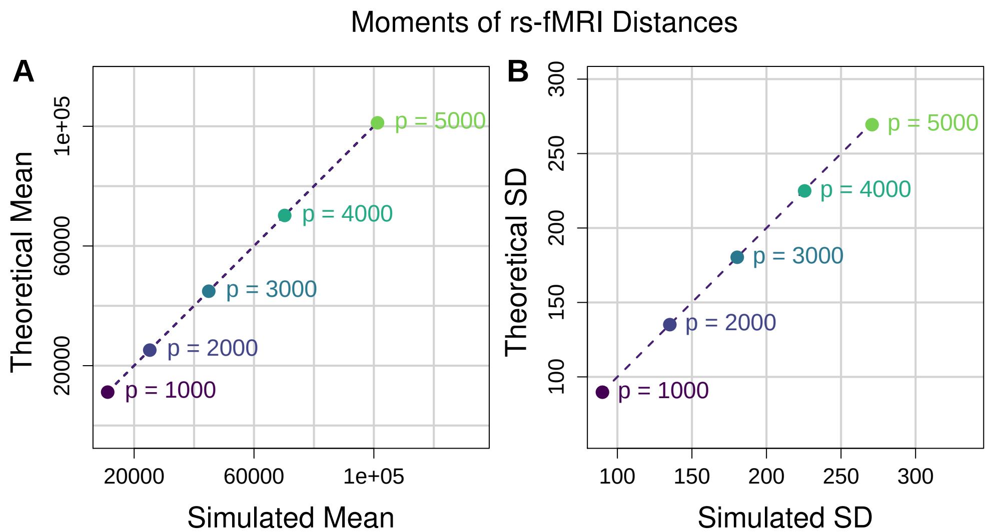

Figure S19. This will be a caption.

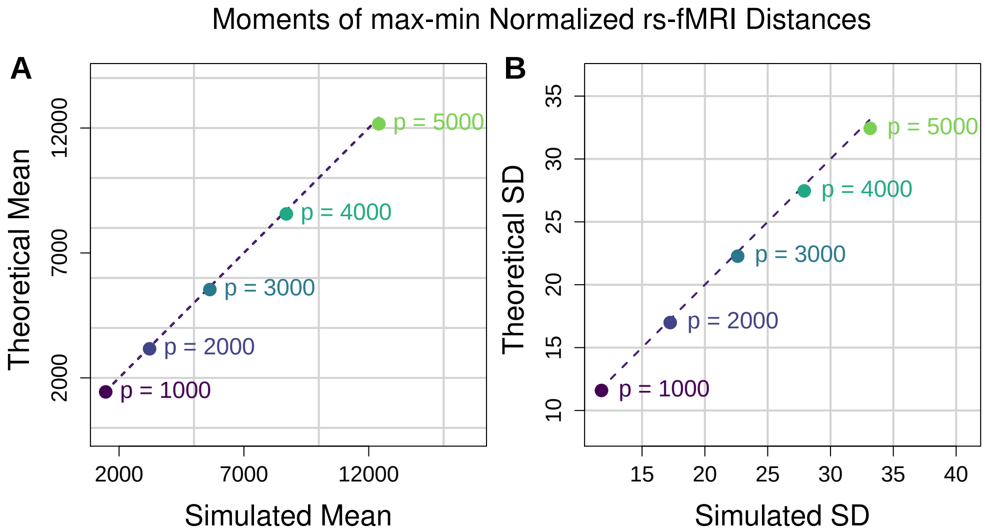

Figure S20. This will be a caption.