Testing the masculinization hypothesis in a sample of 23,935 human brains

Nitay Alon[1], Isaac Meilijson[1], Daphna Joel[2,3*]

[1] School of Mathematical Sciences, Tel Aviv University. Tel-Aviv, Israel.
[2] School of Psychological Sciences, Tel-Aviv University. Tel-Aviv, Israel.
[3] Sagol School of Neuroscience, Tel-Aviv University. Tel-Aviv, Israel.
*Correspondence to: Daphna Joel
Email: djoel@tauex.tau.ac.il

Abstract

For over 60 years, the masculinization hypothesis dominates our understanding of sex effects on the brain. According to this view, the male distribution for single brain measures and for the brain as a whole is shifted away from the female distribution. In the last decade this view has been challenged by evidence that sex effects on single brain features may be opposite under different conditions, resulting in brains comprised of unique mosaics of female-typical and male-typical features. Analysis of 289 MRI-derived measures of grey and white matter from 23935 brains revealed only three brain measures for which the masculinization hypothesis was not rejected in favor of the alternative hypothesis that women and men sample from the same two phenotypes. Moreover, at the individual level, sampling was not consistent across brain measures, as some measures were likely sampled from the female-favored phenotype while others were likely sampled from the male-favored phenotype. Last, considering the relations between brain measures, the brain architecture of women and men was remarkably similar. These results do not support the masculinization hypothesis but are consistent with the mosaic hypothesis as well as with other lines of evidence showing that the brain architectures typical of women are also typical of men, and vice versa, and that sex category explains a very small part of the variability in human brain structure.

Introduction

In spite of conflicting evidence (recently reviewed in 1), the 60 years old masculinization hypothesis still dominates our understanding of sex effects on the brain. According to this hypothesis, sex-related factors masculinize specific features of the brains of males away from a default female form (for review see 2, 3). McCarthy had further suggested that sex effects on the brain are "subject to canalization to assure that males and females are robustly different on multiple end points, but to also assure not too much" (4, p.4). In line with these views, animal and human studies often reveal group-level differences between females and males on macroscopic and microscopic measures of the brain (e.g., brain region volume, number of neurons, 1, 5). Furthermore, in humans, although there is overlap between the distributions of females and males for all currently known brain measures which show a sex difference, the distribution of one sex often seems to be shifted compared to the distribution of the other sex (e.g., 5, 6).

Studies in laboratory animals reveal, however, that a change in environmental conditions (e.g., group housing instead of individual housing) may result in a reversal of the phenotypes typical of females and males (e.g., in layer III of the visual cortex, the dendritic morphology typical of males housed individually was typical of females housed in groups, and the dendritic morphology typical of females housed individually was typical of males housed in groups, *7*; for review of similar findings, see *8*). Such studies further show that a manipulation (e.g., chronic stress) that may reverse sex differences in one brain measure (e.g., density of CB1 receptors in the dorsal hippocampus) may have a different effect on sex differences in another brain measure (e.g., density of CB1 receptors in the ventral hippocampus, where the stress led to the disappearance of a sex difference, *9*).

Keeping in mind the canalization hypothesis, such observations suggest that females and males are not canalized into the 'female' and 'male' phenotypes of a brain feature, but rather that females and males differ in their probability of manifesting each of a feature's two phenotypes. These observations further suggest that the probabilities of manifesting each phenotype may be affected by environmental factors, and that these effects may be different for different brain features. As a result of the latter, different features within a single brain may not be internally consistent with regard to phenotype – always sampling from the phenotype more common in females or always sampling from the phenotype more common in males. Instead, some features will have the phenotype more common in females, while other features will have the phenotype more common in males – the 'mosaic' hypothesis (*8, 10-12*). Using the density of CB1 receptors as an example, in a colony of rats kept under standard laboratory conditions, males would have high receptor density in the dorsal and ventral hippocampus, and females would have low receptor density in the two hippocampal regions (*9*). However, females in this colony that experienced chronic stress (e.g., because of being housed with a dominant rat), would show in the dorsal hippocampus the phenotype typical of males in that colony (high receptor density), whereas in the ventral hippocampus they would exhibit the phenotype typical of females in that colony (low receptor density).

Building on the hypothesis that specific features in the brains of males are masculinized away from a default female form and McCarthy's canalization

hypothesis (*4*), the assumption underlying the present study was that two phenotypes (distributions) underlie the observed distributions of brain measures, and the aim was to discover the relations between these underlying phenotypes and sex category (female, male). We analyzed 289 MRI-derived measures of both grey matter (volume) and white matter (mean diffusivity [MD] and fractional anisotropy [FA]) in the brains of 12,466 women and 11,469 men obtained from the UK Biobank (*13*). Because these measures are correlated with total brain volume (e.g., *14-16*), all analyses were conducted with brain volume taken into account using the power method (*14, 17*). For each brain measure, the expectation-maximization algorithm (*18*) was used to compute maximum likelihood estimators of the parameters (mean and variance) of each of a measure's two underlying distributions as well as the proportions of men and women who sample[1] from each of the distributions. We then tested whether women and men sample from different phenotypes ('pure-types' hypothesis, Fig. 1A) or from the same two phenotypes ('mixed-types' hypothesis, Fig. 1B-E); whether they sample from the two phenotypes with different probabilities (Fig. 1C-D); and whether this sampling is consistent across brain features (e.g., always sampling from the male-favored phenotype). (Note that the analyses reported here were performed under the working assumption that the underlying distributions are Gaussian. We explain in the Supplementary Materials that our approach yields essentially the same answers under a much broader class of distributions, and provide data regarding the suitability of the Gaussian assumption for the data analyzed here).

Results

Out of the 289 brain measures analyzed, in 282 the pure-types hypothesis was rejected – that is, women and men sampled from the two underlying distributions in positive proportions.

Brain measures better described by a pure-types model

The seven measures for which the pure-types hypothesis was not rejected are listed in Table S1. In four of these measures, the sex difference in the observed data was trivial (|Cohen's d| < 0.04; and in three of these also not statistically significant). It therefore seems safe to conclude that each of these four measures is best described as reflecting a single phenotype. In the remaining three pure-types measures – the volumes of: the anterior division of the right cingulate gyrus, the right planum polare, and the left postcentral gyrus – there was a (small) sex difference in the observed data (0.17 < |Cohen's d| < 0.28), suggesting a shift in the distribution of one sex compared to the other, in line with the masculinization hypothesis. The observed distributions of women and men for the measure showing the largest sex difference (the volume of the left postcentral gyrus, Cohen's d = 0.273) are presented in Figure 1A.

[1] References to 'sample' here and elsewhere in the text are used with the following meaning: Each phenotype contains a range of possible values which occur with different frequencies, and the specific value of a brain measure for an individual brain is sampled from this distribution. This value may be sampled from one of two phenotypes, and the major question of the present study concerns the relations between a person's sex category (female, male) and the phenotype from which the value of each of their brain measures is sampled. 'Sample' as used in this paper does not imply an intentional process on the part of brains or humans.

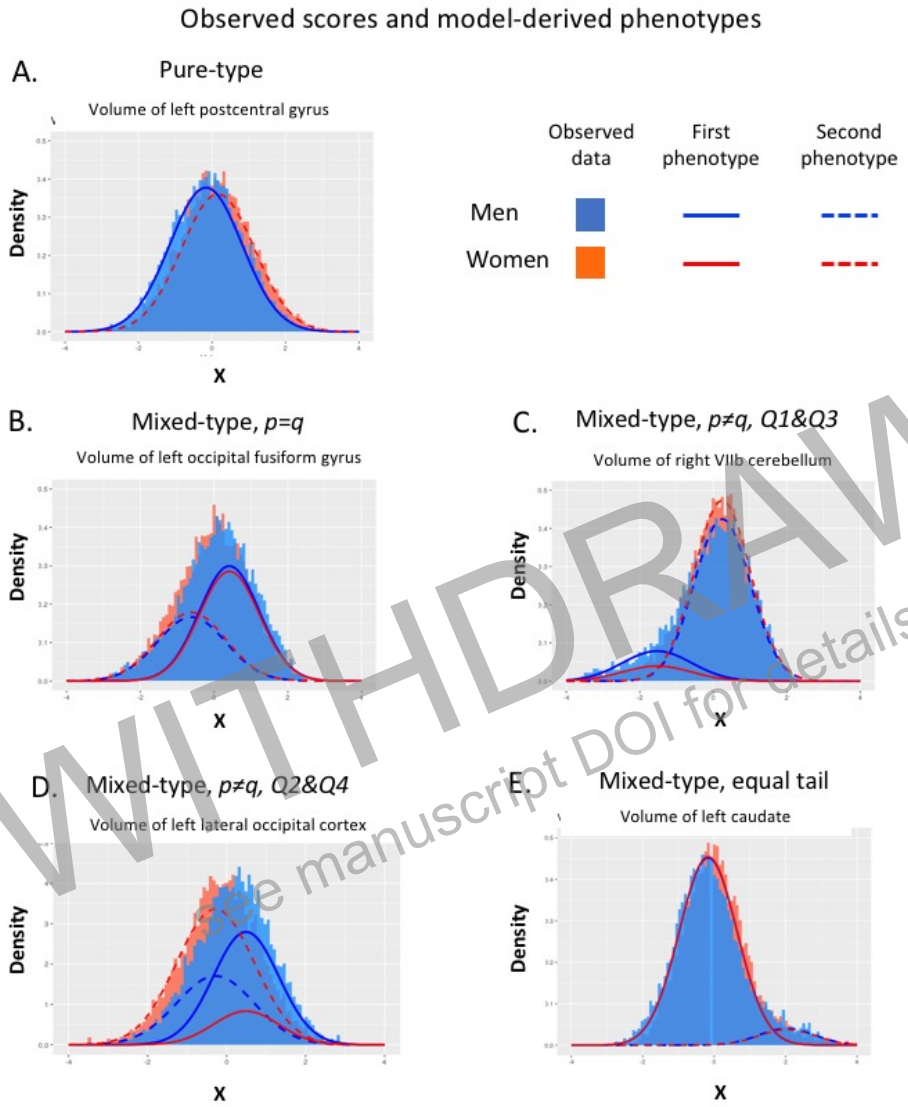

Figure 1. Observed scores and model-derived phenotypes. (A-E) Frequency distributions of the observed scores of women (orange) and men (light blue) for specific brain measures, and of the scores of women (red) and men (blue) on the model-fitted underlying phenotypes - one presented with a dashed line and the other with a solid line. (A) An example of a pure-types measure. Of the seven measures for which the pure-types hypothesis was not rejected, the volume of the left postcentral gyrus showed the largest sex difference in the observed data (Cohen's d = 0.273). (B) An example of a mixed-types measure in which men and women sample with the same probability from the two model-fitted phenotypes ($p = q$). (C) An example of a mixed-types measure in which both men and women 'favor' the same model-fitted phenotype, but with significantly different probabilities ($p \neq q$, Q1&Q3). (D) An example of a mixed-types measure in which men 'favor' one model-fitted phenotype and women 'favor' the other ($p \neq q$, Q2&Q4). Of the 41 such measures, the volume of the left lateral occipital cortex showed the largest sex difference in sampling probabilities (Cohen's h = 0.764) and in the observed data (Cohen's d = 0.338). (E) An example of a mixed-types measure with a "tail" – only a small proportion of humans sample from one of the model-fitted phenotypes. In this example, men and women sample with the same probability from the two model-fitted phenotypes ($p = q$).

Brain measures better described by a mixed-types model

Figure 2 displays for each of the 282 brain measures that were better described by a mixed-types model than by a pure-types model the probability that a man (p, X axis) and a woman (q, Y axis) will sample from the distribution with the higher mean. In 84 of the mixed-types measures, men and women sampled from the two distributions with the same probabilities ($p = q$, e.g., Fig. 1B, 1E), whereas in the remaining 198

measures (marked with a plus symbol in Fig. 2), women and men sampled with significantly different probabilities ($p \neq q$, e.g., Fig. 1C, 1D). In 104 measures, the sex difference in sampling probabilities was small (|Cohen's h| < 0.2), in 86 moderate (0.2 < |Cohen's h| < 0.5), and in eight large (0.5 < |Cohen's h| < 0.765; Figure 1D presents the distributions of women and men for the volume of the left lateral occipital cortex, which showed the largest sex difference in sampling probabilities (Cohen's h = 0.764) and in means in the observed data (Cohen's d = 0.338)). To further appreciate the magnitude of the sex differences in sampling probability, we compared the likelihood that a woman and a man would sample from the same phenotype of a brain measure to the likelihood that two women or two men would sample from the same phenotype. The ratio between the first likelihood and the other two was, on average, 0.979 (range, 0.699-1.170). For comparison, for the 84 $p=q$ measures, the corresponding ratio was, on average, 0.998 (range, 0.902-1.078).

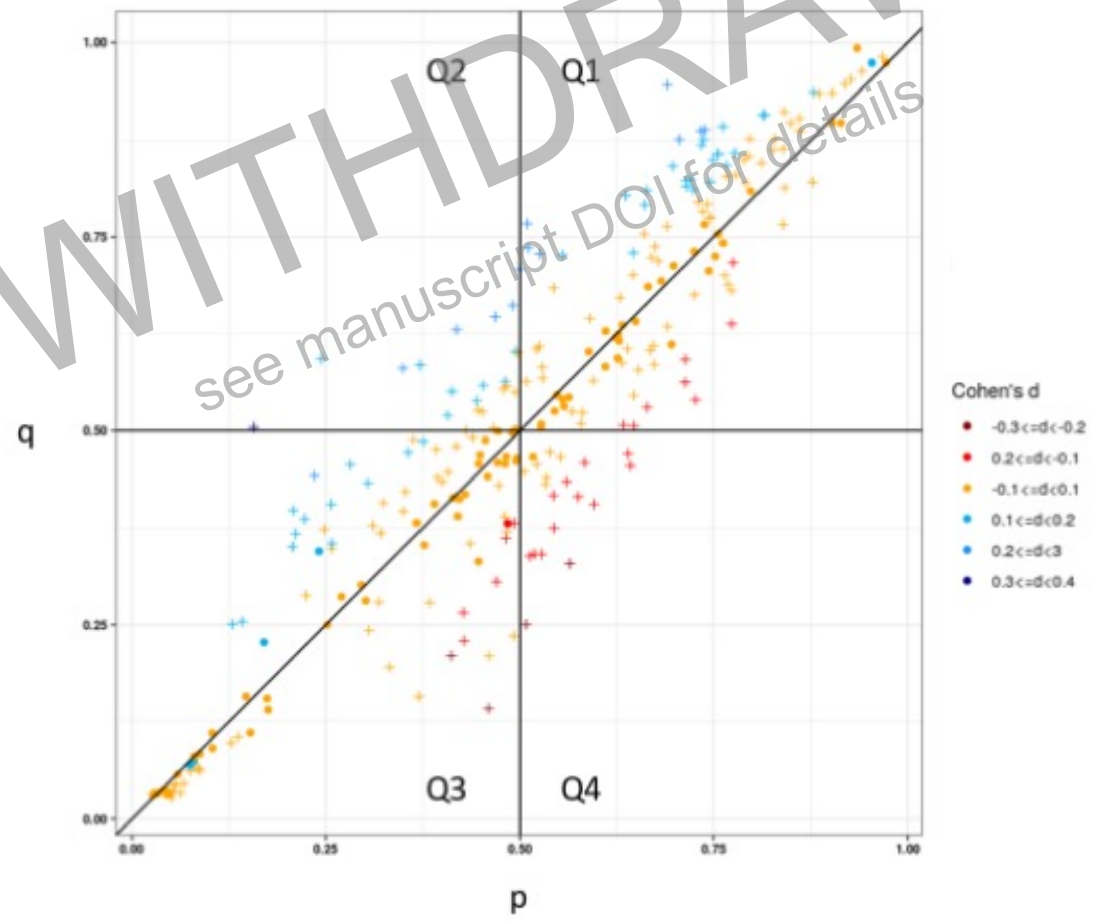

Figure 2. The probability that a man (*p,* X axis) and a woman (*q,* Y axis) will sample from the high-mean distribution of the 282 mixed-types measures. The color-code represents the effect size (Cohen's d) of the sex difference in the observed data; Measures for which there was a significant sex difference in the probability of sampling are marked with a plus symbol.

Is sampling consistent across brain measures?

To test whether different measures within a brain are consistent or mosaic in the phenotype from which they sample, we chose the 41 $p \neq q$ mixed-types features for which one phenotype was sampled mainly by women ('female-favored' phenotype) *and* the other mainly by men ('male-favored' phenotype). These measures are located at Quadrants 2 and 4 of the graph in Fig. 2, and listed in Table S2. Figure 1D presents

the distributions of women and men for one such measure, which showed the largest difference in sampling probabilities. For every participant and for each of these measures the posterior probability[2] that this measure was sampled from the 'male-favored' phenotype was calculated. Then the correlation coefficients between these probabilities were assessed, separately for women and for men, for all possible pairs of same-type measures (i.e., regional volume, mean FA, and mean MD; There were no such measures among the weighted-mean FA and weighted-mean MD). If brains are consistent in the phenotype from which they sample, then high positive correlations are expected between all pairs of measures, whereas if brains are 'mosaics' – each brain with a unique combination of features, some in the 'male-favored' phenotype and some in the 'female-favored' phenotype - then most correlations are expected to be low.

Figures 3a-c present these correlations in women (lower triangle) and men (upper triangle), separately for regional volume, mean FA, and mean MD measures. Correlation strength is represented using a red (-1) – white – blue (+1) color scale, and the absolute size is represented by the size of the dot. Measures of volume were largely uncorrelated, indicating that with respect to regional volume, brains are not consistent in sampling from the male-favored phenotype. In contrast, the correlation coefficients between FA and MD measures occupied a wider range, with both positive and negative moderate correlations. Negative correlations reflect a situation in which sampling the male-favored phenotype in one region correlates with sampling the female-favored phenotype of another region. To better understand these unexpected negative correlations, we assessed the correlations in the same set of measures, but this time between the posterior probabilities to select from the high-mean distribution (Fig. 3D-F). This analysis resulted in positive correlations only (as was also the case in the correlations between the posterior probabilities to select from the high-mean distribution for all other measures of mean FA and mean MD; data not shown), indicating that value (high versus low) is more important than sex category in explaining variability in FA and MD even for measures for which the majority of men sampled from one phenotype and the majority of women sampled from the other. This may also be true for the few moderate-to-large positive correlations between measures of volume (Fig. 3A), which are positive also in Fig. 3D – these correlations may reflect not the consistent effects of sex but rather the consistent effects of other factors. Indeed, the strongest correlations were found between homologous regions in the two hemispheres (the right and left hippocampus; the right and left cuneal cortex; and the right and left superior frontal gyrus).

The most remarkable observation over the images presented in Figure 3 is that the correlation matrices in men and women are almost identical (the numerical values of the correlations are given in Figure S4). This remarkable similarity is also evident when considering the correlations between all possible same-type pairs of the 282 mixed-types measures (Fig. 4. Note that the correlations here are between the posterior probabilities of each individual to select from the high-mean distribution, rather than from the male-favored distribution). The correlations in men (blue) are

[2] 'Posterior probability' refers to the probability that this measure was sampled from the 'male-favored' phenotype, given the observed value of this brain measure in this brain and the underlying distributions parameters.

sorted from lowest to highest, and the correlations in women (red) are presented in the men's order.

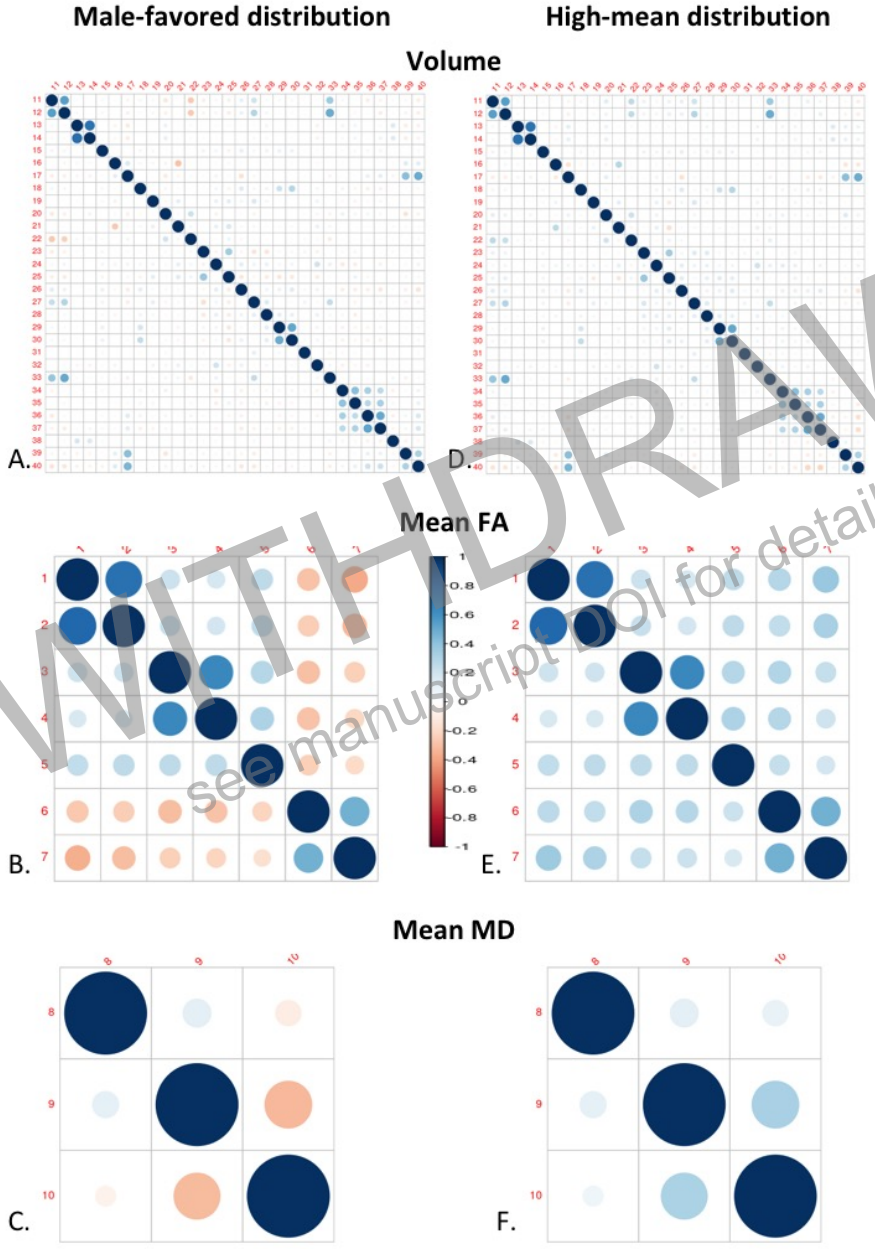

Figure 3. Assessing internal consistency. (A-C) The correlation coefficients in women (lower triangle) and men (upper triangle) between the posterior probabilities of each individual to select from the male-favored distribution in all possible pairs of the 41 measures that have a female-favored and a male-favored distribution, separately for three types of brain measures - (A) volume, (B) mean FA, and (C) mean MD. Correlation strength is represented using a red (-1) – white – blue (+1) color scale, and the absolute size is also represented by the size of the dot. The numbers correspond to the number of each measure in Table S2. (D-F) Same as A-C but for the posterior probabilities of each individual to select from the high-mean distribution.

Discussion

The present analysis does not support the 60 years old assumption that the brains of males are masculinized away from a default female form. Of the 289 brain measures

analyzed, such a description was appropriate for only three measures in which a small shift was evident in the distribution of one sex compared to the distribution of the other sex. For 282 brain measures, the hypothesis that women and men sample from different phenotypes was rejected in favor of the hypothesis that men and women sample from the same two phenotypes. This suggests that if brain measures are described as reflecting two underlying Gaussian-shaped phenotypes (rather than, for example, one non-Gaussian-shaped phenotype, or three Gaussian-shaped phenotypes), then women and men sample from both phenotypes, and for the most part do so with quite similar probabilities. The overall small sex differences in sampling probabilities would contradict also a "soft" version of the masculinization hypothesis, if this existed, according to which the large majority of men sample from one phenotype whereas the large majority of women sample from the other phenotype.

The conclusion that brains of human males are not masculinized away from a default female form was further supported by the pattern of correlations between the posterior probabilities of each individual to select from the male-favored distribution. These correlations were either mainly around zero (for measures of volume), indicating that sampling from the male-favored phenotype for one region provided no information on whether the male-favored or the female-favored phenotype of another region was sampled, or in a pattern suggesting the existence of factor(s) that are more important than sex category in explaining variability in brain measures.

The possibility that sex category is not a major predictor of variability in human brain structure is further supported by the almost identical correlations in women and men between the posterior probabilities to select from the high-mean distribution for all possible pairs of same-type brain measures (Fig. 4). This remarkable similarity suggests that the same principles are governing brain architecture in women and men.

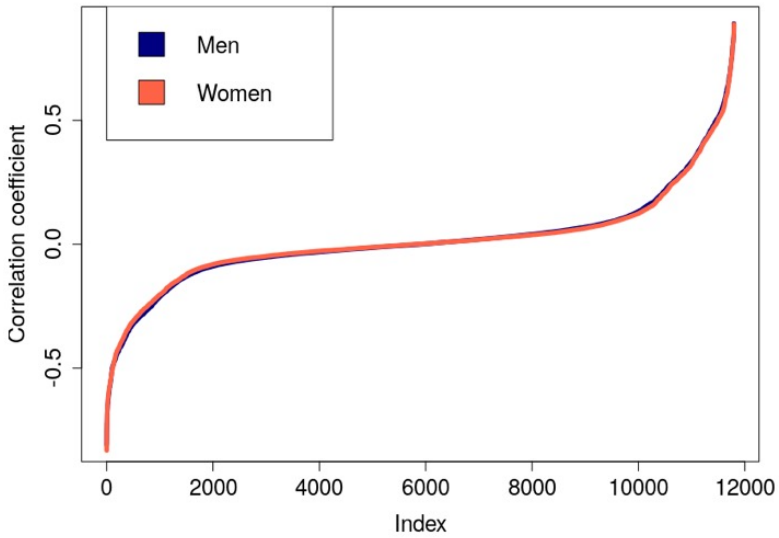

Figure 4. The correlation coefficients in women (red) and men (blue) between the posterior probabilities of each individual to select from the high-mean distribution in all possible same-type pairs of the 282 mixed-types measures. The correlations in men are sorted from lowest to highest, and the correlations in women are presented in the men's order.

Our findings add to other lines of evidence revealing that sex category is not a major predictor of variability in human brain structure and function. Thus, on the basis of a

review of studies of sex differences in human brain structure, Eliot concluded that when brain size is controlled for, sex category accounts for less than 2% of the variability in brain structure (*19*). Recent studies, which assessed the contribution of several factors to variability in brain function (measured using functional MRI), reported that sex category explained only a small fraction of this variability (*20, 21*). Finally, an assessment of the relations between the number of sex differences in functional MRI studies and sample size did not reveal the positive correlation expected if brain function of women and men belonged to two populations (*22*).

We would like to stress that the conclusion that sex category is not a major predictor of variability in human brain structure does not contradict evidence that sex-related genes and hormones affect specific brain measures (for a recent review see *1*), nor evidence that supervised machine learning algorithms may use sex-related variability in brain structure to predict the sex category of a brain's owner (e.g., *23-26*). Indeed, one such approach (logistic regression, as in *23*) over the 289 brain measures analyzed in the present study, accurately predicted the sex category of brains' owners in 75% of cases (see Supplemental Materials; The lower classification rate compared to previous studies (e.g., *23-25*) is expected given that we used data "corrected" for total brain size (*26*)).

What our present and previous studies (*1, 6, 27*) challenge is the common assumption (e.g., *23, 24, 28*) that sex-related effects consistently add up in individual brains so that the brains of women are meaningfully different from the brains of men. Different analytical approaches repeatedly demonstrate that this is not the case. Thus, animal studies show that sex effects interact in complex ways with multiple other factors, to create multimorphic, rather than dimorphic, brains (reviewed in, *8, 10*). In humans, an analysis of internal consistency in MRI-derived brain measures that show large sex differences revealed that mosaic brains (consisting of some measures with scores more common in women compared to men, and some measures with scores more common in men compared to women) are much more prevalent than internally consistent brains (*6*). Mosaic brains were also observed when analyzing post mortem-derived hypothalamic measures which show very large sex differences (*1*). In addition, unsupervised machine learning algorithms applied to the entire brain revealed that the brain architectures typical of women are also typical of men and vice versa; large sex differences were found only in the prevalence of some rare brain architectures (*27*). The present analysis adds to these the observations that for the vast majority of brain measures, women and men sample from the same phenotypes, often with very similar probabilities, and that the same factors are similarly accounting for co-variability of brain measures.

The present study has several limitations. The sample is quite ethnically homogeneous (92.2% Caucasians) and restricted age-wise (all participants are over 42 years old), restricting the generalizability of our conclusions across ethnicity and age. On the other hand, this relative homogeneity would have increased the chances of finding consistent sex effects, if these were present, as other studies have shown that sex differences in brain structure may differ across age (e.g., *29, 30*) and across countries differing in their ethnicity composition (e.g., *6, 31*). Another limitation of the present study is that we analyzed only MRI-derived brain measures, which show smaller sex differences compared to some post mortem-derived measures (e.g., number of neurons in specific hypothalamic nuclei). It is unlikely, however, that a

large enough dataset of the latter type of measures would be available to enable the analyses conducted here.

Conclusions

The decades old hypothesis according to which brains of males are masculinized away from a default female form should be replaced with a more complex model according to which sex-related variables are a part of a large set of factors which similarly interact in women and men to create a highly heterogenous population of human brains (*6, 11, 12, 27*). This population cannot be meaningfully divided into 'female' and 'male' types nor aligned along a female-male continuum (*8, 11*). There is therefore a need to develop new methods for studying the human brain and its relations with sex-related variables that go beyond the common practice of comparing a group of females to a group of males (*12, 27*).

Materials and Methods

Data collection and preparation for analysis

The present study was conducted as part of UK Biobank application 42111, and made use of imaging-derived measures generated by an image-processing pipeline developed and run on behalf of UK Biobank (*32*). Data were derived from the brains of 12,466 women (mean age = 65.16 years, SD = 7.28) and 11,469 men (mean age = 66.52 years, SD = 7.56). The following measures were analyzed: the volume of 139 regions of grey matter; the mean diffusivity (MD) and fractional anisotropy (FA) of 48 tracts defined using the Tract-Based Spatial Statistics analysis; and the weighted-mean MD and FA of a set of 27 major tracts, derived using probabilistic tractography-based analysis (for details of the acquisition protocols, image processing pipeline, and derived measures, see https://biobank.ctsu.ox.ac.uk/crystal/crystal/docs/brain_mri.pdf). All analyses were conducted with brain volume taken into account using the power method (*14, 17*). Total intracranial volume (TIV) was calculated as the sum of the following two variables from the UK Biobank dataset: volume of grey and white matter and volume of ventricular cerebrospinal fluid. For each brain measure, a linear regression of the log of its value versus log-TIV was fitted, and the residuals were used in all subsequent analyses.

Statistical analysis

Expectation-Maximization Algorithm (EM)

The EM algorithm (*18*) is applied to data assumed to be generated from a mixture of parametric distributions with unknown parameters and unknown mixture probabilities. The working assumption is that the data are sampled from two Gaussian distributions, one with parameters, μ_1, σ_1, and the other with parameters, μ_2, σ_2. The proportions of men and women who sample from the high-mean distribution are p and q, respectively. The EM method is used to compute maximum likelihood estimators (MLE) for both the distribution parameters (μ_1, σ_1, μ_2, σ_2) and the proportions (p and q).

In all analyses, the p-value was computed using Wilks' theorem (*33, 34*) and adjusted for FWER using the Benjamini-Hochberg correction (*35*). Adjusted p-values smaller than 0.05 were considered statistically significant.

A power analysis (described in the Supplementary Materials) confirmed that the size of the sample in the present study (12,466 women and 11,469 men) is large enough for the aims of the present study (Figure S1).

Hypothesis testing

Pure-types or mixed-types?

For each brain measure, the null hypothesis is that the pair p and q is either (0,1) or (1,0) - that is, that women sample from a 'female' distribution and men sample from a 'male' distribution. The alternative hypothesis is that there exist two latent distributions with parameters (μ_1,σ_1,and μ_2,σ_2), from which men and women sample with proportions that are larger than 0 and smaller than 1. Equations 1 and 2 describe the mixed-types model: Letting N stand for normal density, p stand for the proportion of men sampling from the high-mean distribution, and q stand for the proportion of women sampling from the high-mean distribution, the density of men's features (X) and women's features (Y) is

$$(1)\ f_X(x,\Theta) = pN(x;\mu_1,\sigma_1) + (1-p)N(x;\mu_2,\sigma_2)$$
$$(2)\ f_Y(x,\Theta) = qN(x;\mu_1,\sigma_1) + (1-q)N(x;\mu_2,\sigma_2)$$

For each brain measure, the MLE of these parameters was estimated by the EM method (see Supplementary Materials). A log-likelihood ratio test was conducted to test the null hypothesis of pure-types versus the alternative hypothesis of mixed-types.

Is there a sex difference in the probability of sampling from the two phenotypes?

Measures for which the pure-types hypothesis was rejected are those for which the observed data belong to a non-Gaussian distribution best described by a mixture of two Gaussians that are sampled by both women and men. To test whether men and women differ in their probabilities of sampling from the two Gaussians, a similar analysis was conducted, but this time the null hypothesis was that $p = q$, and the alternative hypothesis was that $p \neq q$. For brain features for which p was significantly different from q, the size of the difference was estimated using Cohen's h (*36*).

Assessing internal consistency.

Correlation matrices between the EM responsibilities (i.e., the posterior probabilities of each individual to sample from a reference distribution given his/her sex category, *37*) of two sets of measures were evaluated: I. The posterior probabilities to select from the high-mean distribution for all mixed-types measures; II. The posterior probabilities to select from the male-favored distribution for the 41 mixed-types measures for which $p \neq q$ and the female-favored distribution was different from the male-favored distribution (e.g., Fig. 1D). In both sets, Pearson correlation coefficients were computed only between measures of the same type (i.e., separately for measures of volume, mean FA, weighted mean FA, mean MD, and weighted mean MD).

Code.

The data in this work were analyzed using the R programming language (*38*). The equal probability EM was computed using the mixtools package (*39*). Cohen's d was estimated using the effsize package (*40*). The code for the paper is available at https://github.com/nitayalon/biobank_data_analysis.

Acknowledgments

This work has been conducted using the UK Biobank Resource under Application 42111, and supported by the Israel Science Foundation (grant No. 217/16 to DJ and IM).

Author Contributions

DJ framed the research question. NA and IM adapted existing analytic tools and developed new ones. NA ran all computations. DJ, NA and IM wrote the paper.

References

1. D. Joel, A. Garcia-Falgueras, D. Swaab, The Complex Relationships between Sex and the Brain. *Neuroscientist*, 1073858419867298 (2020).
2. M. M. McCarthy, in *Developmental Neuroendocrinology,* S. Wray, S. Blackshaw, Eds. (Springer Nature Switzerland, 2020), pp. 393-412.
3. M. M. McCarthy, A. P. Arnold, Reframing sexual differentiation of the brain. *Nat Neurosci* **14**, 677-683 (2011).
4. D. Joel, M. M. McCarthy, Incorporating Sex As a Biological Variable in Neuropsychiatric Research: Where Are We Now and Where Should We Be? *Neuropsychopharmacology* **42**, 379-385 (2017).
5. S. J. Ritchie *et al.*, Sex Differences in the Adult Human Brain: Evidence from 5216 UK Biobank Participants. *Cereb Cortex* **28**, 2959-2975 (2018).
6. D. Joel *et al.*, Sex beyond the genitalia: The human brain mosaic. *P Natl Acad Sci USA* **112**, 15468-15473 (2015).
7. J. M. Juraska, J. M. Fitch, C. Henderson, N. Rivers, Sex differences in the dendritic branching of dentate granule cells following differential experience. *Brain Res* **333**, 73-80 (1985).
8. D. Joel, Male or Female? Brains are Intersex. *Front Integr Neurosci* **5**, 57 (2011).
9. C. G. Reich, M. E. Taylor, M. M. McCarthy, Differential effects of chronic unpredictable stress on hippocampal CB1 receptors in male and female rats. *Behav Brain Res* **203**, 264-269 (2009).
10. D. Joel, Genetic-gonadal-genitals sex (3G-sex) and the misconception of brain and gender, or, why 3G-males and 3G-females have intersex brain and intersex gender. *Biol Sex Differ* **3**, 27 (2012).
11. D. Joel, in *Sex Differences in Neurology and Psychiatry,* R. Lanzenberger, G. S. Kranz, I. Savic, Eds. (Elsevier BV, San Diego, 2020), vol. 175, pp. 13-24.
12. D. Joel, A. Fausto-Sterling, Beyond sex differences: new approaches for thinking about variation in brain structure and function. *Philos T R Soc B* **371**, 20150451 (2016).
13. K. L. Miller *et al.*, Multimodal population brain imaging in the UK Biobank prospective epidemiological study. *Nat Neurosci* **19**, 1523-1536 (2016).
14. C. Sanchis-Segura *et al.*, Sex differences in gray matter volume: how many and how large are they really? *Biol Sex Differ* **10**, 32 (2019).

15. H. Takao, N. Hayashi, S. Inano, K. Ohtomo, Effect of head size on diffusion tensor imaging. *Neuroimage* **57**, 958-967 (2011).
16. S. B. Vos, D. K. Jones, M. A. Viergever, A. Leemans, Partial volume effect as a hidden covariate in DTI analyses. *Neuroimage* **55**, 1566-1576 (2011).
17. D. Liu, H. J. Johnson, J. D. Long, V. A. Magnotta, J. S. Paulsen, The power-proportion method for intracranial volume correction in volumetric imaging analysis. *Front Neurosci* **8**, 356 (2014).
18. A. P. Dempster, N. M. Laird, D. B. Rubin, Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)* **39**, 1–38 (1997).
19. L. Eliot, in *Cambridge International Handbook on Psychology of Women*. F. Cheung, D. Halpern, Eds. (Cambridge University Press, Cambridge, UK, 2020).
20. A. J. Kersey, K. D. Csumitta, J. F. Cantlon, Gender similarities in the brain during mathematics development. *NPJ Sci Learn* **4**, 19 (2019).
21. E. Mitricheva, R. Kimura, N. K. Logothetis, H. R. Noori, Neural substrates of sexual arousal are not sex dependent. *Proc Natl Acad Sci U S A* **116**, 15671-15676 (2019).
22. S. P. David *et al.*, Potential Reporting Bias in Neuroimaging Studies of Sex Differences. *Sci Rep* **8**, 6082 (2018).
23. A. M. Chekroud, E. J. Ward, M. D. Rosenberg, A. J. Holmes, Patterns in the human brain mosaic discriminate males from females. *P Natl Acad Sci USA* **113**, E1968-E1968 (2016).
24. M. Del Giudice *et al.*, Joel et al.'s method systematically fails to detect large, consistent sex differences. *P Natl Acad Sci USA* **113**, E1965-E1965 (2016).
25. D. Joel, A. Persico, J. Hanggi, J. Pool, Z. Berman, Reply to Del Giudice Et Al., Chekroud Et Al., and Rosenblatt: Do Brains of Females and Males Belong to Two Distinct Populations? *P Natl Acad Sci USA* **113**, E1969-E1970 (2016).
26. C. Sanchis-Segura, M. V. Ibanez-Gual, N. Aguirre, A. J. Gomez-Cruz, C. Forn, Effects of different intracranial volume correction methods on univariate sex differences in grey matter volume and multivariate sex prediction. *Sci Rep* **10**, 12953 (2020).
27. D. Joel *et al.*, Analysis of Human Brain Structure Reveals that the Brain "Types" Typical of Males Are Also Typical of Females, and Vice Versa. *Front Hum Neurosci* **12**, 399 (2018).
28. M. Ingalhalikar *et al.*, Sex differences in the structural connectome of the human brain. *Proc Natl Acad Sci U S A* **111**, 823-828 (2014).
29. L. Jancke, S. Merillat, F. Liem, J. Hanggi, Brain Size, Sex, and the Aging Brain. *Hum Brain Mapp* **36**, 150-169 (2015).
30. R. K. Lenroot, J. N. Giedd, Sex differences in the adolescent brain. *Brain Cogn* **72**, 46-55 (2010).
31. K. Zilles, R. Kawashima, A. Dabringhaus, H. Fukuda, T. Schormann, Hemispheric shape of European and Japanese brains: 3-D MRI analysis of intersubject variability, ethnical, and gender differences. *Neuroimage* **13**, 262-271 (2001).
32. F. Alfaro-Almagro *et al.*, Image processing and Quality Control for the first 10,000 brain imaging datasets from UK Biobank. *Neuroimage* **166**, 400-424 (2018).
33. A. W. van der Vaart, *Asymptotic Statistics*. (Cambridge University Press, Cambridge, 1998).
34. S. S. Wilks, The large-sample distribution of the likelihood ratio for testing composite hypotheses. *Ann. Math. Statist.* **9**, 60- 62 (1938).

35. Y. Benjamini, Y. Hochberg, Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society: Series B (Methodological)* **57**, 289-300 (1995).
36. J. Cohen, *Statistical Power Analysis for the Behavioral Sciences*. (Lawrence Erlbaum Associates, Hillsdale, N.J., ed. 2nd, 1988).
37. T. Hastie, R. Tibshirani, J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction.* (Springer, ed. 2nd, 2001).
38. R Core Team, R: A language and environment for statistical computing. R Foundation for Statistical Computing (Vienna, Austria, 2013). URL http://www.R-project.org/.
39. T. Benaglia, D. Chauveau, D. R. Hunter, D. Young, mixtools: An R Package for Analyzing Finite Mixture Models. *Journal of Statistical Software* **32**, 1-29 (2009).
40. M. Torchiano, _effsize: Efficient Effect Size Computation. R package version 0.8.0 (2020). URL: https://doi.org/10.5281/zenodo.1480624