Supplemental Materials and Methods

Permissions: This project was approved by the Danish Twin Registry Board (approved 10/11/2007), and the research ethics boards of the IWK Health Centre and the Capital District Health Authority, Halifax, Canada (Project 4040).

Data collection and treatment: Individual age and XIP data for two adult populations were available from two previous studies (Kristiansen 2005, Sandovici 2004). For both, XIPs were determined by the differential methylation of the human androgen receptor (HUMARA) in DNA extracted from peripheral blood as described elsewhere (Allen 1992). The HUMARA assay is considered the gold standard in research and clinical laboratories (Ørstavik 2009, Amos-Landgraf 2006). The reliability of this assay and the relationship between HUMARA locus methylation and X-inactivation status have been confirmed (Amos-Landgraf 2006, Busque 2009), as has the biological relevance of the assay (reviewed in Ørstavik 2009 .) The present study examines two datasets. The *Utah* dataset comprises age and XIP data from 183 individuals reported as 'time 1' in the original study (Sandovici 2004). The *Denmark* dataset consists of 258 informative 'twin 1' individuals from the dizygotic pairs included in the original study (Kristiansen 2005). For both datasets, care was taken at the time of data collection to ensure accurate XIP determination, such as performing multiple technical replicates for each individual female and by ensuring enzyme efficiency by testing control samples, such as male DNA, which is not subject to XCI-induced methylation and must therefore be completely digested during the assay (Kristiansen 2005, Sandovici 2004).

In order to thoroughly and rigorously analyze our datasets, a new data treatment strategy was developed. First, XIP data are folded about XIP = 0.5 (Sup Fig 1). Folded-XIPs are equivalent to the relative abundance of cells from the smaller cell population, be that the maternal-X or paternal-X population. Folded-XIPs fall within the range of 0.00 to 0.50 where 0.00 indicates extreme skewing and 0.50 a perfectly balanced folded-XIP. Data folding masks parent of origin effects, but allows for the inclusion of all data, without the need to arbitrarily assign phase when this is unknown. It does not bias the distribution in any way. Others have used similar strategies for example, by looking at the prevalence of the most prevalent population (Bolduc 2008), or the absolute value of the shift away from a balanced XIP (Wong 2011). Alternatively, some have pooled two groups together in their final analysis, for example by considering females with patterns > 0.20 and < 0.80 to be "skewed" (Kristiansen 2005). Next, data are represented as empirical cumulative distribution functions (CDF), rather than histograms (Sup Fig 1). This avoids arbitrarily binning data and generates a curve

which is convenient for testing empirical data against continuous statistical models using goodness of fit tests.

Formulation of statistical models

Models of completely random choice: If XCI choice is random and independent in each cell, then the possible primary XIPs are i/c, for $i = 0,\ldots,c$, where c is the number of progenitor cells undergoing choice, and the probability of inactivating the maternal X is equal to the probability of inactivating the paternal X. The distribution of the possible primary XIPs in a population therefore follows a scaled binomial (c, 0.5) distribution with mean $1/2$ and standard deviation $\frac{1}{\sqrt{4c}}$ (Fig 2a). It is expected that XIPs will shift from their primary value established at the time of choice to some related secondary value due, for example, to cell loss, developmental bottlenecks, or proliferation differences between maternal and paternal X populations. The distribution of XIPs may then transform from a discrete function to a continuous one. Three methods are used to model continuous distributions of secondary XIPs (Sup Fig 2a-c). For the "simple normal model", the discrete binomial distribution described above is approximated by a normal distribution with the same mean and variance (Sup Fig 2b, solid line). This model has been used elsewhere (Fialkow 1973). For the "simple beta model", the discrete distribution is approximated by a beta distribution, with probability restricted to the permissible [0,1] interval (Sup Fig 2b, dashed line). Various integer values for c, the number of progenitor cells, are tested for each simple model. To find the best possible fitting simple random models, c is estimated from the data allowing for any real number. For the simple normal model, a maximum likelihood estimate is obtained, taking the data folding into account, using the relationship: $\hat{c} = n/(4\sum_{i=1}^{n}(y_i - 0.5)^2)$ where n is the sample size and y_i, $i = 1,\ldots, n$ are the folded-XIP values. A standard error for the estimate can be obtained using the nonparametric bootstrap. To estimate the number of progenitor cells for the simple beta model, we note that the beta distribution is symmetric when the two parameters, α and β, are equal. The variance of this distribution is $1/4(2\alpha + 1)$. Equating this to the scaled binomial variance $1/4c$ gives the relationship $c = 2\alpha + 1$, so an estimate for α can be converted to an estimate of c. There is no explicit maximum likelihood estimate for α, so numerical techniques are used. For the "mixed beta model", secondary XIPs for individuals with primary XIP equal to i/c, for $i = 1,\ldots,c$ are assumed to follow a beta distribution centered at $m_i = i/c$ (Sup Fig 2c, dashed lines), where i is the number of cells with the paternal X active. The variances of these secondary skewing distributions are assumed to decrease with

increased primary skewing according to the equation: $\sigma_i^2 = \tau^2 m_i (1 - m_i)$, where τ^2 is a multiplicative factor to be estimated from the data. A consequence of this choice is that the distributions corresponding to m_i equal to 0 or 1 have zero variance, and so a very small fraction, $1/2^7 = 0.0078$, of individuals with these extreme values are assumed not to be affected by secondary skewing. The overall distribution of XIPs is a probability weighted sum of the component distributions, with weights given by the binomial probabilities from the model for primary skewing (Sup Fig 2c, solid line). The beta distribution is skewed left for $m_i < 0.5$ and skewed right for $m_i > 0.5$, and we place the modes of the components at the primary XIP values. Numerical methods are used to solve for the beta parameters α and β given the mode m_i and variance σ_i^2. To ensure the component beta distributions are unimodal, the variance scaling factor τ^2 must be less than 0.577.

Models of genetically influenced choice: Statistical models of genetically influenced human XCI choice were created using the basic tenants of the mouse *Xce* paradigm in order to test whether empirical data is consistent with the idea of a human *XCE*. These models suppose that a single locus with two or more alleles influences XIPs. A mixture model assigns primary XIP values of 0.5 for homozygotes and $0.5 \pm \delta_{ij}$ for heterozygotes with alleles *i* and *j*, with skewing of $\delta_{ij} > 0$ toward the parent contributing the dominating allele (Sup Fig 2d, f). The probabilities associated with the primary values are: for heterozygotes, the genotype frequencies under the assumption of HWE, $p_i p_j$, and for homozygotes, the total homozygosity $\sum p_i^2$, where p_i is the frequency of the *i*th allele and $\sum p_i = 1$. Genetic models are dependent on the number of allelic variants. Two discrete models were generated for this study: the "2-allele" and the "3-allele" models. For the 2-allele model, the primary values are $0.5 - \delta_{12}$, 0.5, and $0.5 + \delta_{12}$, with associated probabilities $p_1 p_2$, $p_1^2 + p_2^2$ and $p_1 p_2$, with $p_1 + p_2 = 1$ (Sup Fig 2d). For the 3-allele model, the primary values are $0.5 - \delta_{23}$, $0.5 - \delta_{13}$, $0.5 - \delta_{12}$, 0.5, $0.5 + \delta_{12}$, $0.5 + \delta_{13}$ and $0.5 - \delta_{23}$, with $\delta_{12} \leq \delta_{13} \leq \delta_{23}$, and the corresponding probabilities are $p_2 p_3$, $p_1 p_3$, $p_1 p_2$, $p_1^2 + p_2^2 + p_3^2$, $p_1 p_2$, $p_1 p_3$ and $p_2 p_3$, with $p_1 + p_2 + p_3 = 1$ (Sup Fig 2f). Secondary skewing is accommodated using beta distributions centered at the primary skewing values (Sup Fig 2e, g). The same variance structure is employed here as was used for the mixed beta models of completely random XCI. The skewing coefficients δ_{ij}, allele frequencies p_i and variance multiplicative factor τ^2 are estimated from the data using the method of maximum

likelihood and computer iteration. The R language (*A language and environment for statistical computing*) is used for this and all other statistical calculations. Both the 2- and 3-allele genetic models described here are independent of progenitor cell number. The models are also independent of the location of the assumed gene; our genetic models could accommodate a gene located at the proposed human *XCE* locus, Xq25, or elsewhere on the X-chromosome, such as within the *XIC* as is the case for the mouse *Xce* (Simmler 1993), or even on an autosome. We do not include the effects of strong negative selections such as X:autosomal translocations, X-linked lethal alleles, or dominant *XIST* mutations which lead to complete XIP skewing since these are considered here to be very rare in a healthy female population. In support of this assumption, we find no individual in either dataset with a folded XIP of 0. The genetic models described here do not consider multiple independently segregating loci. The development and testing of other genetic models is beyond the scope of this paper.

Application to data: The models described above are symmetric around XIP = 0.5 and the probability above 0.5 is reflected (folded) onto the [0, 0.5] interval to be compared to the folded XIP data.

Statistical comparisons: A two-sample *t* test is used to compare the mean age for *Utah* and *Denmark* samples. Lowess curves and linear regression are used to determine whether there is a significant correlation between age and XIP. The Anderson-Darling statistic (D'Agostino 1986) is used to test for goodness-of-fit of the models (one sample), and to compare the distributions of folded-XIP for the *Utah* and *Denmark* samples (two sample). A customized form of the statistic was derived for models with a mass of probability at 0, as these distributions are neither continuous nor discrete (Pritchett 2009). Significance levels for the Anderson-Darling statistic are obtained using the parametric bootstrap, where a reference distribution is constructed by repeated sampling from the fitted model (Davidson 1997). Comparisons between models are based on likelihood ratio statistics. For nested models, significance is assessed using the appropriate χ^2 distribution; otherwise a null distribution is derived from 1000 bootstrap samples generated under the simpler model of the null hypothesis. For the comparison of the completely random and 3-allele genetic models, the genetic model is extended to allow a small fraction of folded XIP values at zero. Similarly, a likelihood ratio test is used to determine whether the shifts δ_{ij} are the same for the *Utah* and *Denmark* datasets by comparing the fit of a model which assumes these shifts are the same in

both datasets to a model which uses two sets of shifts.

References:

- Allen RC, Zoghbi HY, Moseley AB, Rosenblatt HM, Belmont JW: Methylation of HpaII and HhaI sites near the polymorphic CAG repeat in the human androgen-receptor gene correlates with X chromosome inactivation. *Am J Hum Genet* 1992. **51**(6): 1229–1239.
- Amos-Landgraf JM, Cottle A, Plenge RM, *et al.*: X chromosome-inactivation patterns of 1,005 phenotypically unaffected females. *Am J Hum Genet* 2006. **79**(3): 493–499.
- Bolduc V, Chagnon P, Provost S, *et al.*: No evidence that skewing of X chromosome inactivation patterns is transmitted to offspring in humans. *J Clin Invest* 2008. **118**(1): 333–341.
- Busque L, Paquette Y, Provost S, *et al.*: Skewing of X-inactivation ratios in blood cells of aging women is confirmed by independent methodologies. *Blood* 2009. **113**: 3472–3474.
- D'Agostino R, Stephens M: Goodness-of-Fit techniques. Marcel Dekker, inc., 1986.
- Davidson A, Hinkley D: Bootstrap methods and their application. Cambridge University Press, 1997.
- Fialkow PJ: Primordial cell pool size and lineage relationships of five human cell types. *Ann Hum Genet* 1973. **37**(1): 39–48.
- Kristiansen M, Knudsen GP, Bathum L, *et al.*: Twin study of genetic and aging effects on X chromosome inactivation. *Eur J Hum Genet* 2005. **13**(5): 599–606.
- Ørstavik KH: X chromosome inactivation in clinical practice. *Hum Genet* 2009. **126**: 363–373.
- Pritchett S: Investigating random and genetic models for X-chromosome inactivation. Master's thesis, Dalhousie University, 2009.
- R Development Core Team: R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria, 2010. ISBN 3-900051-07-0.
- Sandovici I, Naumova AK, Leppert M, Linares Y, Sapienza C: A longitudinal study of X-inactivation ratio in human females. *Hum Genet* 2004. **115**(5): 387–392.
- Simmler MC, Cattanach BM, Rasberry C, Rougeulle C, Avner P: Mapping the murine Xce locus with (CA)n repeats. *Mamm Genome* 1993. **4**(9): 523–530.
- Wong CC, Caspi A, Williams B, Houts R, Craig IW, Mill J: A longitudinal twin study

of skewed X chromosome-inactivation. *PLoS ONE* 2011. **6**: e17873.