From correlation to causation: Data-driven exploration of transcriptional regulation using population genomics

Luijk, R.

4 AUTOSOMAL GENETIC VARIATION IS ASSOCIATED WITH DNA METHYLATION IN REGIONS VARIABLY ESCAPING X-CHROMOSOME INACTIVATION

René Luijk, H. Wu, C.K. Ward-Caviness, E. Hannon, E. Carnero-Montoro, J.L. Min, P. Mandaviya, M. Müller-Nurasyid, H. Mei, S.M. van der Maare, BIOS Consortium, C. Relton, J. Mill, M. Waldenberger, J.T. Bell, R. Jansen, A. Zhernakova, L. Franke, P.A.C. 't Hoen, D.I. Boomsma, C.M. van Duijn, M.M.J. van Greevenbroek, J.H. Veldink, C. Wijmenga, J. van Meurs, L. Daxinger, P.E. Slagboom, E.W. van Zwet, B.T. Heijmans

Nature Communications, **9**(1) (2018)

Abstract

The inactivation of one of the female X chromosomes restores equal expression of X-chromosomal genes between females and males. While most of the X-chromosomal genes are silenced by X-chromosome inactivation (XCI), 15% of genes remain bi-allelically expressed, and 10% show variable degrees of escape from XCI between females. However, little is known about genes involved in human XCI, or the causes of variable XCI. Using a discovery data-set of 1,867 females and 1,398 males and an independent replication sample of 3,351 females, we show that genetic variation at three autosomal loci is associated with female-specific changes in X-chromosome methylation. Through cis-eQTL expression analysis in the same 1,867 females, we mapped the loci to the genes *SMCHD1/METTL4*, *TRIM6/HBG2*, and *ZSCAN9*. Low-expression alleles of the loci were predominantly associated with mild hypomethylation of CpG islands near genes known to variably escape XCI, implicating the autosomal genes in variable XCI. Together, these results suggest a genetic basis for variable escape from XCI and highlight the potential of a population genomics approach to identify genes involved in XCI.

Introduction

To achieve dosage equivalency between male and female mammals, one of two X-chromosomes is silenced early in female embryonic development resulting in one inactive (Xi) and one active (Xa) copy of the X-chromosome [Lyon, 1961]. While the Xi-linked gene *XIST* is crucial for the initiation of X-chromosome inactivation (XCI), autosomal genes appear to be critically involved in XCI establishment and maintenance [Galupa and Heard, 2015]. An abundance of repressive histone marks [Brinkman et al., 2006; Heard et al., 2001; Plath et al., 2003] and DNA methylation [Sharp et al., 2011; Yasukochi et al., 2010] throughout XCI on Xi is in line with a prominent role of epigenetic regulation in both phases. However, the Xi is not completely inactivated. With an estimated 15% of X-chromosomal genes consistently escaping XCI, and an additional 10% escaping XCI to varying degrees [Carrel and Willard, 2005; Cotton et al., 2013], escape from XCI is fairly common in humans [Carrel and Willard, 1999; Cotton et al., 2014; Zhang et al., 2013], much more so than in mice [Yang et al., 2010]. Genes escaping XCI are characterized by distinct epigenetic states [Peeters et al., 2014] and are thought to be associated with adverse outcomes, including mental impairment [Peeters et al., 2014; Yang et al., 2010; Zhang et al., 2013].

In the mouse, an example of an autosomal gene involved in XCI is *Smchd1*. *Smchd1* is an epigenetic repressor and plays a critical role in the DNA methylation maintenance of XCI in mice [Blewitt et al., 2008; Nozawa et al., 2013]. However, in humans, in-depth knowledge on the role of autosomal genes in XCI maintenance is lacking, despite earlier *in vitro* efforts [Massah et al., 2014]. Furthermore, the mechanisms underlying variable XCI, a common feature of human XCI [Carrel and Willard, 2005; Cotton et al., 2013], are unknown.

Here, we report on the identification of four autosomal loci associated with female-specific changes in X-chromosome DNA methylation using a discovery set of 1,867 females and 1,398 males, and replication of three of these loci in an independent replication set consisting of 3,351 female samples. The replicated loci map to the genes *SMCHD1/METTL4*, *TRIM6/HBG2* and *ZSCAN9* through eQTL analysis. All three preferentially influenced the methylation of CpGs located in CpG islands near genes known to variably escape XCI between individuals, providing evidence for a genetic basis of this phenomenon.

Results

Identification of female-specific genetic effects on X-chromosome methylation

To identify genetic variants involved in XCI, we employed a global test approach [Luijk et al., 2015] to evaluate the association of 7,545,443 autosomal genetic variants with DNA methylation at any of 10,286 X-chromosomal CpGs measured in whole blood of 1,867 females (Supplementary Data 1) using the Illumina 450k array (see Methods). The analysis was corrected for covariates, including

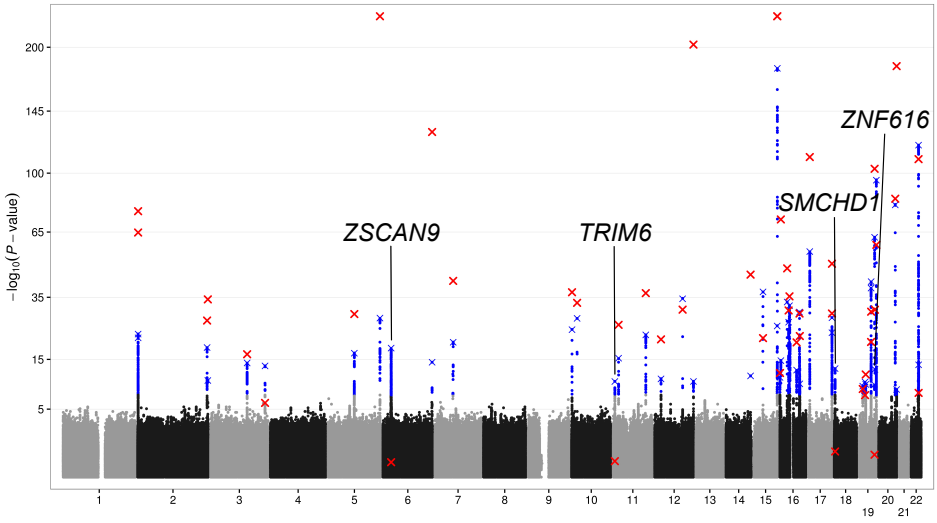

Figure 4.1: Manhattan plot showing all tested autosomal SNPs for an overall effect on X-chromosomal methylation in females. Significant associations are depicted in blue (Wald $P < 5 \times 10^{-8}$). The sentinel variant per independent locus is indicated with a blue cross. Testing the effects of these 48 sentinel variants in males, we found 44 replicated in males (Wald $P < 1.1 \times 10^{-6}$, red cross), whereas the other 4 loci were female-specific, as they clearly lacked an effect in males (Wald $P > 0.19$).

cell counts, age, and batch effects. We identified 4,504 individual variants representing 48 independent loci associated with X-chromosomal methylation in females (Wald $P < 5 \times 10^{-8}$, Figure 4.1 and Supplementary Fig. 1), each defined by the most strongly associated variant (as reflected by the lowest P-value), termed the sentinel variant. Of the 48 sentinel variants corresponding to these 48 loci, 44 were also associated with X-chromosomal methylation in males ($N = 1,398$, Supplementary Data 1, Supplementary Data 2; Wald $P < 1.1x10^{-6}$) indicating that the associations were unrelated to XCI. The four remaining variants did not show any indication for an effect in males (Wald $P > 0.19$) while they did show strong, widespread, and consistent same-direction effects across the X-chromosome in females (Supplementary Data 2, Supplementary Data 3, Supplementary Fig. 2). Formally testing for a genotype by sex interaction revealed significant interaction effects for three of the four variants. The rs140837774, rs139916287, and rs1736891 variants with evidence for an interaction ($P_{interaction} < 5.9 \times 10^{-4}$) mapped to the *SMCHD1/METTL4*, *TRIM6/HBG2*, and *ZSCAN9* loci, respectively, (see Methods). The remaining variant rs73937272 ($P_{interaction} = 0.88$) mapped near the *ZNF616* gene. Finally, we evaluated whether the effect of the autosomal loci was influenced by genetic variation on X, but this did not change the results (Supplementary Data 4, see Methods).

To establish the validity and stability of the analyses, we first investigated whether any of the associations were due to confounding by cellular heterogeneity. Therefore, we directly tested for an association between the four identified sentinel variants and the observed red and white blood cell counts. This did not result in any significant association (Supplementary Fig. 3). Furthermore, we determined that none of the four identified variants are among the variants known to affect blood composition [Orru et al., 2013; Roederer et al., 2015]. Vice versa, genetic variants known to affect blood cell counts also did not show an association with X-chromosomal methylation in our data (Supplementary Fig. 4, Supplementary Data 5). Re-testing the effects of the four sentinel variants while adjusting for nearby blood composition-associated SNPs (< 1Mb) did not influence the results (Supplementary Data 6, see Methods). Second, we addressed unknown confounding by including latent factors as covariates in our models, estimated in the methylation data using software for estimation and adjustment of unknown confounders in high-dimensional data (Wang et al. [2015], see Methods). Re-testing the four sentinel variants without these latent factors did not change the results (Supplementary Data 6). We conclude that the effects of the four variants identified in the discovery data are stable and not confounded by cellular heterogeneity or other, unknown, factors.

Finally, we tested the four sentinel variants in an additional 3,351 unrelated female samples (see Methods and Supplementary Data 7), and successfully replicated the rs140837774, rs139916287, and rs1736891 variants (Bonferroni corrected, $P_{adj} = 0.0096$, $P_{adj} = 2.4 \times 10^{-4}$, and $P_{adj} = 2.2 \times 10^{-3}$, respectively). The rs73937272 variant ($P_{adj} = 1$), which also lacked a sex-genotype interaction effect in the discovery set, was not replicated. In further analyses, we focussed on the three replicated loci.

Exploration of genetic loci with female-specific effects on X-methylation

The sentinel variant rs140837774 is an AATTG insertion/deletion variant (MAF = 0.49) on chromosome 18, located in intron 26 of *SMCHD1* (Supplementary Fig. 2), a gene known to be critically involved in XCI in mice [Blewitt et al., 2008; Chen et al., 2015; Daxinger et al., 2013; Gendrel et al., 2012, 2013]. In addition, *SMCHD1* mutations affect the methylation levels of the *D4Z4* repeat in humans, playing an important role in facioscapulohumeral dystrophy 2 (FSHD2, Lemmers et al. [2012]). To link rs140837774 to a nearby gene we performed a *cis*-eQTL analysis using RNA-seq data from the 1,867 females in the discovery set of our study (250Kb up- and downstream of the sentinel variant, Supplementary Data 8). We found that the deletion was strongly associated with decreased *SMCHD1* expression (Fisher's $P = 1.8 \times 10^{-10}$, regression coefficient = -0.13) and increased expression of the methyltransferase *METTL4*, albeit weaker (Wald $P = 4.9 \times 10^{-4}$, regression coefficient = 0.04). *METTL4* is a highly conserved gene [Breiling and Lyko, 2015; Falckenhayn et al., 2016], involved in the mRNA modification N^6-methyladenosine (reviewed in Gilbert et al. [2016]), which plays an important role in epigenetic regulation in mammals [Wu et al., 2016].

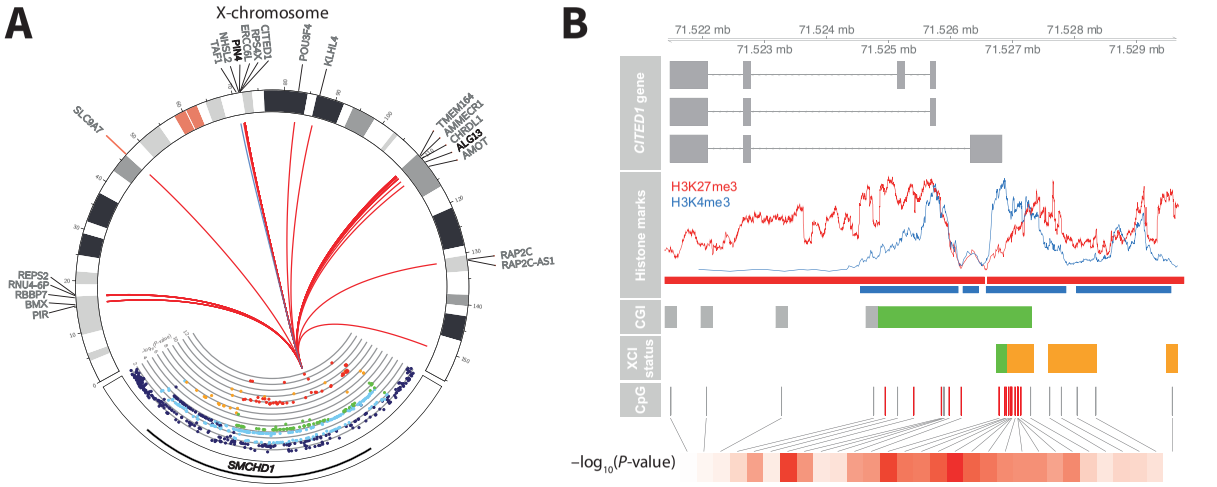

Figure 4.2: The *SMCHD1/METTL4* locus associates with DNA methylation at X-chromosomal and autosomal regions. **(A)** Plot showing the *SMCHD1/METTL4* locus and the effects it has on the X-chromosome. The colors in the *SMCHD1/METTL4* locus indicate LD (red: $R^2 \geq 0.8$; orange: $0.6 \leq R^2 < 0.8$; green: $0.4 \leq R^2 < 0.6$; light blue: $0.2 \leq R^2 < 0.4$; dark blue: $R^2 \leq 0.6 \leq 0.2$). The y-axis shows the –$\log_{10}$(P-value) of the association with overall X-chromosomal methylation. The line colors in the Circos plot indicate the direction of the effect (red: hypomethylation, blue: hypermethylation). The plot of the *SMCHD1/METTL4* locus mostly shows moderate to high LD and covers most of the *SMCHD1* gene. The deletion of rs140837774 is associated with both downregulation of *SMCHD1* and upregulation of *METTL4* (see main text). The deletion of rs140837774 is associated with hypomethylation at 98.2% of all associated CpGs (56 CpGs, red lines, mean effect size 1% per allele). Hypomethylation of CpGs near two genes known to escape XCI to varying degrees (PIN4 and ALG13, shown in bold) is associated with increased expression of these two genes. **(B)** Example of CpG island (CGI) in the CITED1 gene (first row) associated with the *SMCHD1/METTL4* locus. The CpGs associated with this locus (fifth row, indicated by red lines) are overrepresented in regions characterized by both active (blue) and repressive (red) histone marks (second row, red and blue bars, 2-fold enrichment, Fisher's $P = 1.5 \times 10^{-12}$; 6.9-fold enrichment, Fisher's $P = 1.5 \times 10^{-12}$), are often located in CpG islands (third row, green bar, 11.3-fold enrichment, Fisher's $P = 2.5 \times 10^{-14}$), and regions known to variably escape X-chromosome inactivation (fourth row, orange bars, 21.4-fold enrichment, Fisher's $P = 3.7 \times 10^{-18}$). The bottom row indicates the strength of the associations in terms of –$\log_{10}$(P-value) (dark red indicates strong associations).

The *SMCHD1/METTL4* variant was associated with altered methylation levels of 57 X-chromosomal CpGs in females (FDR < 0.05, Figure 4.2A, Supplementary Data 9). The deletion (the low *SMCHD1* expression allele) was associated with hypomethylation at 56 of those X-chromosomal CpGs (98.2%, binomial $P = 8.5 \times 10^{-13}$ Figure 4.2A), consistent with X-hypomethylation in female mice deficient for *SMCHD1* [Gendrel et al., 2013]. The mean effect size was 1% per rare allele (ranging from 0.27% to 2.34%, Supplementary Fig. 5), with the mean methylation values per CpG ranging from 2.6% to 55% (average methylation at these 56 CpGs is 23.6%).

Compared to all X-chromosomal CpGs in our data, the associated X chromosomal CpGs were strongly overrepresented in CpG islands (50 out of 57 CpGs, 11.3-fold enrichment, binomial $P = 2.5 \times 10^{-14}$, Figure 4.2B), in line with *SMCHD1*'s role in X-chromosomal CpG island methylation [Gendrel et al., 2012]. Data on chromatin marks in blood (Kundaje et al. [2015], see Methods) revealed a strong overrepresentation of the associated X-chromosomal CpGs in regions bivalently marked by the active histone mark H3K4me3 (47 CpGs, 8.2-fold enrichment, Fisher's $P = 1.5 \times 10^{-12}$), and the repressive mark H3K27me3 (38 CpGs, 6.9-fold enrichment, Fisher's $P = 1.5 \times 10^{-12}$), as compared to all X-chromosomal CpGs in our data. In agreement with this, we observed a 16.9-fold enrichment for CpGs overlapping bivalent/poised transcription start sites (TSSs) (35/57 CpGs, Fisher's $P = 4.3 \times 10^{-23}$) using predicted chromatin segmentations [Kundaje et al., 2015], possibly reflecting the mixed signals from both the active and inactive X chromosomes underlying these chromatin segmentations. Strikingly, annotation by the degree of escape for 489 TSSs in 27 different tissues, and specifically whole blood (Cotton et al. [2014], see Methods), revealed a strong overrepresentation of CpGs located near TSSs variably escaping XCI (22 CpGs, 21.4-fold enrichment, Fisher's $P = 3.7 \times 10^{-18}$, Figure 4.2B). Only a modest enrichment for associated CpGs in fully escaping XCI regions (15 CpGs, 4.2-fold enrichment, Fisher's $P = 4.5 \times 10^{-5}$) and an underrepresentation of associated CpGs in inactivated regions (7 CpGs, 28.6-fold depletion, Fisher's $P = 2.2 \times 10^{-23}$) was observed.

Further supporting a link with variable XCI, we observed that X-chromosomal CpGs were associated with differential expression of the nearby genes (<250Kb, see Methods) *ALG13* and *PIN4* (see Methods, Supplementary Data 10) both known to variably escape XCI [Zhang et al., 2013]. While a strong eQTL effect and a clear biologically relevant link with XCI mainly implies *SMCHD1* in X-chromosomal hypomethylation (insertion of rs140837774), an eQTL effect for *METTL4*, although slightly weaker, leaves open a possible role for *METTL4* in XCI, given its role in the mRNA modification N^6-methyladenosine.

Using both female and male samples ($N = 3,265$, Supplementary Fig. 6) to investigate associations of genetic variation at the *SMCHD1/METTL4* locus with autosomal methylation in trans (>5Mb), we found that the *SMCHD1/METTL4* variant was associated with 20 CpGs mapping to the *HOXD10*, *HOXC10*, and *HOXC11* genes of the HOXD and HOXC clusters located on chromosomes 2 and 12, as well as to the large protocadherin beta (PCDHβ) and gamma (PCDHγ) clusters on chromosome 5 (FDR < 0.05, Supplementary Fig. 7, Supplementary

Data 11), all known *SMCHD1* targets [Gendrel et al., 2013; Mould et al., 2013].

The second of the three sentinel variants, sentinel SNP rs139916287 (MAF = 0.07), is located in intron 4 of the *HBG2* gene on chromosome 11, in the β-globin locus (rs139916287, chromosome 11, Supplementary Fig. 2B). The rare allele of the sentinel variant was associated with decreased expression of both the *HBG2* and *TRIM6* genes (T allele; Wald $P = 5.3 \times 10^{-7}$, regression coefficient = -130.55; Wald $P = 9.8 \times 10^{-5}$, regression coefficient = -0.05; Supplementary Data 8), based on *cis*-eQTL mapping testing genes up to 250kb up-, and downstream of the sentinel variant (Supplementary Data 8). While *HBG2* showed higher expression levels and a stronger eQTL effect in our data, *TRIM6* has been shown to bind *XIST* [Chu et al., 2018], and contributes to the maintenance of pluripotency in mouse embryonic stem cells [Sato et al., 2012], making *TRIM6* a strong candidate for explaining our observations. Associating the *TRIM6/HBG2* variant with X-chromosomal methylation, we found 276 associated X-chromosomal CpG sites (FDR < 0.05, Figure 4.3A, Supplementary Data 9). The rare allele (T allele) was associated with hypomethylation at 258 of those CpGs (93.5%, binomial $P = 6.3 \times 10^{-47}$), where mean effect size at these 258 CpGs is 1.6% per T allele, ranging from 0.15% to 4.25% (Supplementary Fig. 5).

Similar to the *SMCHD1/METTL4* variant, associated CpGs were overrepresented in CGIs (199 CpGs, 4.2-fold enrichment, Fisher's $P = 1.6 \times 10^{-29}$), and enriched in regions characterized by H3K27me3 (208 CpGs, 10.5-fold enrichment, Fisher's $P = 3.5 \times 10^{-77}$) and H3K4me3 (Kundaje et al. [2015], 217 CpGs, 6.4-fold enrichment, Fisher's $P = 5.5 \times 10^{-46}$, Figure 4.3B). The associated CpGs were again strongly overrepresented in genomic regions variably escaping XCI in an external set of whole blood samples (Cotton et al. [2014], see Methods, 39 CpGs, 8.8-fold enrichment, Fisher's $P = 2.1 \times 10^{-20}$), to a lesser extent present in regions consistently escaping XCI (61 CpGs, 6.8-fold enrichment, Fisher's $P = 2.2 \times 10^{-23}$), and underrepresented in repressed regions (39 CpGs, 15.2-fold depletion, Fisher's $P = 1.9 \times 10^{-50}$).

In addition, many genes known to variably escape XCI [Zhang et al., 2013], annotated to CpGs associated with *TRIM6/HBG2* genetic variation (*ALG13*, *ATP6AP2*, *CXorf38*, *MED14*, *SMC1A*, *TBL1X*, Supplementary Data 10). Similar to *SMCHD1/METTL4*, these results suggest a role for the *TRIM6/HBG2* locus in variable escape from XCI.

The sentinel SNP of the third identified locus (rs1736891, MAF = 0.38) was associated with the expression of several nearby genes annotated as zinc fingers (Supplementary Data 8), but most strongly with downregulation of the expression of the nearby transcription factor *ZSCAN9* gene [Vaquerizas et al., 2009] located on chromosome 6, based on *cis*-eQTL mapping in our own data (Wald $P = 2.5 \times 10^{-49}$, Supplementary Data 8). The sentinel SNP was significantly associated with 19 X-chromosomal CpGs in females (FDR < 0.05, Supplementary Fig. 8, Supplementary Data 9), all located in the same CpG island: the high-expression A allele of rs1736891 was associated with mild hypomethylation of all 19 CpGs (Fisher's $P = 3.6 \times 10^{-5}$, mean effect size 1.3% per allele, Supplementary Fig.

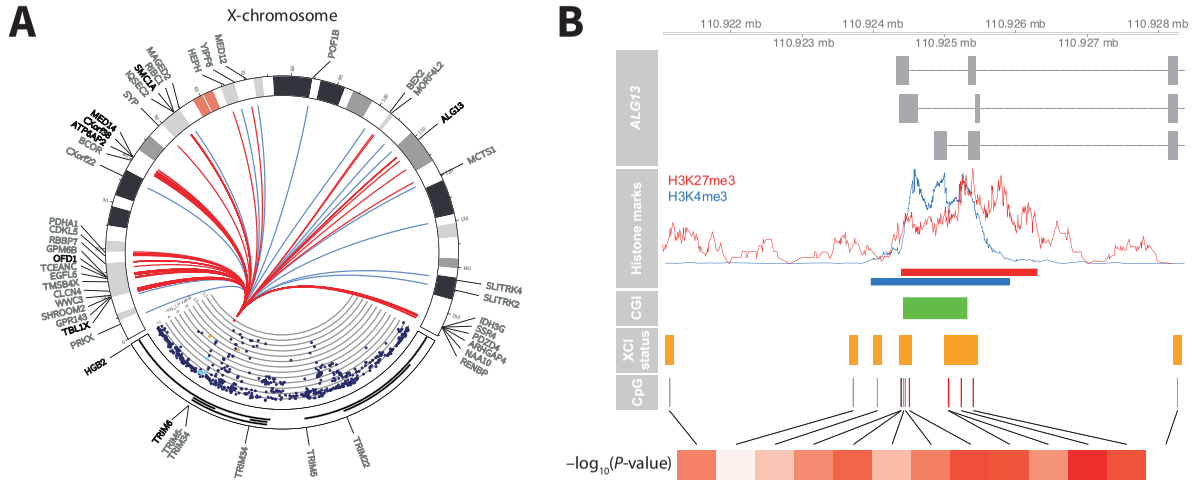

Figure 4.3: The *TRIM6/HBG2* locus is associated with DNA methylation at X-chromosomal regions. **(A)** Plot showing the *TRIM6/HBG2* locus and the widespread effects it has on the X-chromosome. The colors in the *TRIM6/HBG2* locus indicate LD (red: $R^2 \geq 0.8$; orange: $0.6 \leq R^2 < 0.8$; green: $0.4 \leq R^2 < 0.6$; light blue: $0.2 \leq R^2 < 0.4$; dark blue: $R^2 \leq 0.6 \leq 0.2$). The y-axis shows the $-\log_{10}(P\text{-value})$ of the association with overall X-chromosomal methylation. The line colors in the Circos plot indicate the direction of the effect (red: hypomethylation, blue: hypermethylation). The T allele of its sentinel variant rs139916287 is associated with upregulation of *HBG2*, downregulation of *TRIM6* (both shown in bold), and hypomethylation at 258 of the 276 associated CpGs (93.5%, red lines, mean effect size 1.6% per allele). X-chromosomal genes whose expression levels were associated with methylation levels of nearby CpGs are shown in bold. **(B)** Example of CpG island (CGI) in the *ALG13* gene associated with the *TRIM6/HBG2* locus. The enrichments of CpGs in certain genomic regions are similar to those found for the *SMCHD1/METTL4* locus. Most notably, the associated CpGs are also overrepresented in regions known to variably escape X-chromosome inactivation (fourth row, orange bars, 8.8-fold enrichment, Fisher's $P = 2.1 \times 10^{-20}$).

5). There was an overlap of in the CpGs associated with the sentinel variants of the *ZSCAN9* and the *SMCHD1/METTL4* locus, although the two loci are located on different chromosomes (chromosomes 6 and 18, respectively, Supplementary Fig. 2). These associations were statistically independent from each other (*i.e.*, additive), as all identified loci were identified using conditional analyses (see Methods). Specifically, 17 out of 19 CpGs (89.5%) were also targeted by the *SMCHD1/METTL4* locus, and all 19 CpGs show consistent opposite effects for both loci (Supplementary Fig. 9). Similar to the *SMCHD1/METTL4* locus, the *ZSCAN9* locus also associated with autosomal DNA methylation in trans (>5Mb). However, none of the autosomal CpGs overlapped between the two loci (Supplementary Data 11).

Discussion

Here, we identify three autosomal genetic loci with female-specific effects on X-chromosomal methylation in humans (*SMCHD1/METTL4*, *TRIM6/HBG2*, and *ZSCAN9*), all of which were associated with altered expression of autosomal genes *in cis*. Furthermore, all three loci were consistently associated with mild hypomethylation of CpGs overrepresented in CpG islands of X-chromosomal regions known to variably escape XCI in whole blood [Cotton et al., 2014; Zhang et al., 2013]. The former finding extended to 26 other tissues [Cotton et al., 2014], suggesting a cross-tissue genetic basis for variable escape from XCI. We observed a striking underrepresentation of affected CpGs in fully inactivated CGIs, which may be due to the tightly regulated nature of these regions. Methylation of these CpGs may be impervious to the impact of autosomal genetic variation or effects may be substantially weaker requiring much larger data sets to detect them.

While most of the previous work on XCI was done using mouse models and established a critical role for *SMCHD1* in XCI [Blewitt et al., 2008; Daxinger et al., 2013; Gendrel et al., 2013], we here confirm the role of the *SMCHD1/METTL4* locus in XCI in humans and highlight its impact on variable escape from XCI. This phenomenon has not been previously described in mice, perhaps due to the lack of genetic variability in the often inbred mice, leading to less (variable) escape from XCI than occurs in humans [Carrel and Willard, 2005; Cotton et al., 2013]. We also observed associations of the *SMCHD1/METTL4* locus with known autosomal *SMCHD1* targets [Gendrel et al., 2013; Mould et al., 2013], most notably the protocadherin clusters [Mason et al., 2017]. Interestingly, similar to the X-chromosome, the expression of the clustered protocadherin genes is stochastic and mono-allelic [Chess, 2005], suggesting a common mechanism.

In addition to the *SMCHD1/METTL4* locus, our results indicated a role for the *TRIM6/HBG2* locus in XCI. *TRIM6* is a strong candidate to influence female X-chromosome methylation because it was reported to bind *XIST* [Chu et al., 2018] and is involved in *MYC* and *NANOG* regulation [Sato et al., 2012]. Similarly, our data suggest a role for the *ZSCAN9* locus in variable escape from XCI, as it affects a single CpG island that is also targeted by the *SMCHD1/METTL4* locus. While this does suggest a role for the two loci in the same pathway, the effects on the

X-chromosome were statistically independent.

Given the biological consistency of the findings presented here, and the replication thereof in an independent set of samples, our data support a role of autosomal genetic variants in regulating Xi methylation in particular at variably escaping regions. However, to definitely demonstrate causality, unequivocally identify the responsible genes, and provide precise insight into the exact underlying mechanisms, *in vitro* experiments are needed. Importantly, a population genomics approach, like ours, will reveal effects on both XCI establishment and maintenance, which occur during different developmental stages and may involve different molecular pathways. At this point, the exact role of the *SMCHD1/METTL4*, *TRIM6/HBG2* and *ZSCAN9* loci during these processes remain to be determined. Therefore, it will be crucial to design experiments that can discriminate between an effect during the establishment and maintenance phases.

In conclusion, variable escape from XCI in humans has a genetic basis and we identified three autosomal loci, one previous implicated in XCI in mice and two new loci, that influence regions that are susceptible to variable escape from XCI by controlling X-chromosomal DNA methylation or correlated epigenetic marks.

Methods

Discovery cohorts

The Biobank-based Integrative Omics Study (BIOS) Consortium comprises six Dutch biobanks: Cohort on Diabetes and Atherosclerosis Maastricht (CODAM, van Greevenbroek et al. [2011]), LifeLines-DEEP (LLD, Tigchelaar et al. [2015]), Leiden Longevity Study (LLS, Schoenmaker et al. [2005]), Netherlands Twin Registry (NTR, Boomsma et al. [2002]), Rotterdam Study (RS, Hofman et al. [2013]), Prospective ALS Study Netherlands (PAN, Huisman et al. [2011]). The data that were analyzed in this study came from 3,265 unrelated individuals (Supplementary Data 1). Genotype data, DNA methylation data, and gene expression data were measured in whole blood for all samples. In addition, sex, age, measured cell counts (lymphocytes, neutrophils, monocytes, eosinophils, basophils, and red blood cell counts), and information on technical batches were obtained from the contributing cohorts. The Human Genotyping facility (HugeF, Erasmus MC, Rotterdam, The Netherlands, `http://www.glimdna.org`) generated the methylation and RNA-sequencing data and supplied information on technical batches.

Genotype data were generated within each cohort. Details on the genotyping and quality control methods have previously been detailed elsewhere (LLD: Tigchelaar et al. [2015]; LLS: Deelen et al. [2014a]; NTR: Lin et al. [2016]; RS: Hofman et al. [2013]; PAN: Huisman et al. [2011]).

For each cohort, the genotype data were harmonized towards the Genome of the Netherlands (GoNL, The Genome of the Netherlands Consortium et al.

[2014]) using Genotype Harmonizer [Deelen et al., 2014b] and subsequently imputed per cohort using Impute2 [Howie et al., 2009] and the GoNL reference panel (v5, The Genome of the Netherlands Consortium et al. [2014]). We removed SNPs with an imputation info-score below 0.5, a HWE P-value $< 10^{-4}$, a call rate below 95%, or a minor allele frequency smaller than 0.01. These imputation and filtering steps resulted in 7,545,443 SNPs that passed quality control in each of the datasets.

A detailed description regarding generation and processing of the gene expression data can be found elsewhere [Zhernakova et al., 2017]. Briefly, total RNA from whole blood was deprived of globin using Ambion's GLOBIN clear kit and subsequently processed for sequencing using Illumina's Truseq version 2 library preparation kit. Paired-end sequencing of 2x50bp was performed using Illumina's Hiseq2000, pooling 10 samples per lane. Finally, read sets per sample were generated using CASAVA, retaining only reads passing Illumina's Chastity Filter for further processing. Data were generated by the Human Genotyping facility (HugeF) of ErasmusMC (The Netherlands, see URLs). Initial QC was performed using FastQC (v0.10.1), removal of adaptors was performed using cutadapt (v1.1, Martin [2011]), and Sickle (v1.2, Joshi Fass, J. [2011]) was used to trim low quality ends of the reads (minimum length 25, minimum quality 20). The sequencing reads were mapped to human genome (HG19) using STAR (v2.3.0e, Dobin et al. [2013]).

To avoid reference mapping bias, all GoNL SNPs (`http://www.nlgenome.nl/?page_id=9`) with MAF > 0.01 in the reference genome were masked with N. Read pairs with at most 8 mismatches, mapping to as most 5 positions, were used.

Gene expression quantification was determined using base counts (for a detailed description, see Zhernakova et al. [2017]). The gene definitions used for quantification were based on Ensembl version 71. For data analysis, we used reads per kilobase per million mapped reads (RPKM), and only used protein coding genes with sufficient expression levels (median RPKM > 1), resulting in a set of 10,781 genes. To limit the influence of any outliers still present in the data, the data were transformed using a rank-based inverse normal transformation within each cohort.

The Zymo EZ DNA methylation kit (Zymo Research, Irvine, CA, USA) was used to bisulfite-convert 500 ng of genomic DNA, and 4 μl of bisulfite-converted DNA was measured on the Illumina HumanMethylation450 array using the manufacturer's protocol (Illumina, San Diego, CA, USA). Preprocessing and normalization of the data were done as described earlier [Tobi et al., 2015]. Removal of ambiguously mapped probes or probes containing known common genetic variants [Chen et al., 2013] were removed, followed by quality control (QC) using MethylAid's default settings [van Iterson et al., 2014], investigating methylated and unmethylated signal intensities, bisulfite conversion, hybridization, and detection P-values. Filtering of individual beta-values was based on detection P-value ($P < 0.01$), number of beads available (≤ 2) or zero values for signal intensity. Normalization was done using Functional Normalization [Fortin et al., 2014] as implemented

in the minfi R package [Aryee et al., 2014], using five principal components extracted using the control probes for normalization. All samples or probes with more than 5% of their values missing were removed, based on the QC performed using MethylAid. The final dataset consisted of 440,825 probes measured in 3,265 samples. Lastly, similar to the RNA-sequencing data, the methylation data were also transformed using a rank-based inverse normal transformation within each cohort, to limit the influence of any remaining outliers while removing any systematic differences in mean methylation between cohorts.

Replication cohorts

Accessible Resource for Integrative Epigenomics Studies (ARIES) Samples were drawn from the Avon Longitudinal Study of Parents and Children (ALSPAC, Fraser et al. [2013]; Boyd et al. [2013]). Blood from 1022 mother-child pairs (children at three time points and their mothers at two time points) were selected for analysis as part of Accessible Resource for Integrative Epigenomic Studies (ARIES, `http://www.ariesepigenomics.org.uk/`).

Written informed consent has been obtained for all ALSPAC participants. Ethical approval for the study was obtained from the ALSPAC Ethics and Law Committee and the Local Research Ethics Committees. Genotyping and methylation measurements have been previously described [Gaunt et al., 2016; Min et al., 2017].

We are extremely grateful to all the families who took part in the ALSPAC study, the midwives for their help in recruiting them, and the whole ALSPAC team, which includes interviewers, computer and laboratory technicians, clerical workers, research scientists, volunteers, managers, receptionists and nurses.

This work was supported by the UK Medical Research Council; Wellcome (`www.wellcome.ac.uk`; [grant number 102215/2/13/2 to ALSPAC]); the University of Bristol to ALSPAC; the UK Economic and Social Research Council (`www.esrc.ac.uk`; [ES/N000498/1] to CR) and the UK Medical Research Council (`www.mrc.ac.uk`; grant numbers [MC_UU_12013/1, MC_UU_12013/2 to JLM, CR]).

Exeter, Schizophrenia Phase 1 The University College London case-control sample has been described elsewhere [Datta et al., 2008; Hannon et al., 2016] but briefly comprises of unrelated ancestrally matched schizophrenia cases recruited from NHS mental health services and controls from the United Kingdom. Each control subject was interviewed to confirm that they did not have a personal history of an RDC defined mental disorder or a family history of schizophrenia, bipolar disorder, or alcohol dependence. UK National Health Service multicentre and local research ethics approval was obtained and all subjects signed an approved consent form after reading an information sheet. Details of DNA methylation and genetic data generation, processing, quality control and normalisation can be found in the original EWAS manuscript [Hannon et al., 2016].

Exeter, Schizophrenia Phase 2 The Aberdeen case-control sample has been described elsewhere [The International Schizophrenia Consortium, 2008; Hannon et al., 2016] but briefly contains schizophrenia cases and controls who have self-identified as born in the British Isles (95% in Scotland). Controls were volunteers recruited through general practices in Scotland. Practice lists were screened for potentially suitable volunteers by age and sex and by exclusion of subjects with major mental illness or use of neuroleptic medication. Volunteers who replied to a written invitation were interviewed using a short questionnaire to exclude major mental illness in individual themselves and first degree relatives. All cases and controls gave informed consent. The study was approved by both local and multiregional academic ethical committees. Details of DNA methylation and genetic data generation, processing, quality control and normalisation can be found in the original EWAS manuscript [Hannon et al., 2016].

Cooperative health research in the Region of Augsburg Study (KORA F4) The KORA study (Cooperative health research in the Region of Augsburg) consists of independent population-based samples from the general population living in the region of Augsburg, Southern Germany. Written informed consent has been given by each participant and the study was approved by the local ethical committee. The dataset comprised individuals from the KORA F4 survey (all with genotyping and methylation data available) conducted during 2006–2008. The KORA study was initiated and financed by the Helmholtz Zentrum München - German Research Center for Environmental Health, which is funded by the German Federal Ministry of Education and Research (BMBF) and by the State of Bavaria. Furthermore, KORA research has been supported within the Munich Center of Health Sciences (MC-Health), Ludwig-Maximilians-Universität, as part of LMUinnovativ.

Twins UK The TwinsUK cohort was established in 1992 to recruit monozygotic and dizygotic twins [Moayyeri et al., 2013]. More than 80% of participants are healthy female Caucasians (age range from 16 to 98 years old). The cohort includes more than 13,000 twin participants from all regions across the United Kingdom, and many have had multiple visits over the years. The TwinsUK cohort has been used in many epidemiological studies and is representative of the general UK population for a wide range of diseases and traits [Andrew et al., 2001].

TwinsUK is funded by the Wellcome Trust, Medical Research Council, European Union (EU), and the National Institute for Health Research (NIHR)-funded BioResource, Clinical Research Facility, and Biomedical Research Centre based at Guy's and St Thomas' NHS Foundation Trust in partnership with King's College London.

Rotterdam Study The Rotterdam Study (RS) is a large prospective, population-based cohort study aimed at assessing the occurrence of and risk factors for chronic (cardiovascular, endocrine, hepatic, neurological, ophthalmic, psychiatric, dermatological, oncological, and respiratory) diseases in the elderly [Ikram et al., 2017]. The study comprises 14,926 subjects in total, living in the well defined

Ommoord district in the city of Rotterdam in the Netherlands. In 1989, the first cohort, Rotterdam Study-I (RS-I) comprised of 7,983 subjects with age 55 years or above. In 2000, the second cohort, Rotterdam Study-II (RS-II) was included with 3,011 subjects who had reached an age of 55 or over in 2000. In 2006, the third cohort, Rotterdam Study-III (RS-III) was further included with 3,932 subjects with age 45 years and above. Each participant gave an informed consent and the study was approved by the medical ethics committee of the Erasmus University Medical Center, Rotterdam, the Netherlands.

The generation and management of the Illumina 450K methylation array data (EWAS data) for the Rotterdam Study was executed by the Human Genotyping Facility of the Genetic Laboratory of the Department of Internal Medicine, Erasmus MC, the Netherlands. The EWAS data was funded by the Genetic Laboratory of the Department of Internal Medicine, Erasmus MC, and by the Netherlands Organization for Scientific Research (NWO; project number 184021007) and made available as a Rainbow Project (RP3; BIOS) of the Biobanking and Biomolecular Research Infrastructure Netherlands (BBMRI-NL). We thank Mr. Michael Verbiest, Ms. Mila Jhamai, Ms. Sarah Higgins, Mr. Marijn Verkerk, and Lisette Stolk PhD for their help in creating the methylation database. The Rotterdam Study is funded by Erasmus Medical Center and Erasmus University, Rotterdam, Netherlands Organization for the Health Research and Development (ZonMw), the Research Institute for Diseases in the Elderly (RIDE), the Ministry of Education, Culture and Science, the Ministry for Health, Welfare and Sports, the European Commission (DG XII), and the Municipality of Rotterdam. The authors are grateful to the study participants, the staff from the Rotterdam Study and the participating general practitioners and pharmacists.

Identifying female-specific genetic effects influencing X-chromosomal DNA methylation in the discovery cohorts

To identify autosomal genetic variants influencing DNA methylation anywhere on the X-chromosome we applied a two-step approach [Luijk et al., 2015] using 1,867 female samples from the replication cohorts for which both genotype data and methylation data were available. We first fitted linear models to test for an association between each autosomal SNP i and each of 10,286 X-chromosomal CpGs j individually, correcting for known covariates M (cell counts, cohort, age, technical batches - *e.g.*, sample plate and array position) and unknown confounding by including latent factors U, estimated using *cate* [Wang et al., 2015], where the eigenvalue difference method implemented in *cate* suggested an optimal number of three latent factors:

$$y_j = \beta_{ij} x_i + \gamma M + \delta U \tag{4.1}$$

For each autosomal genetic variant i, this approach yields 10,286 P-values p_{ij}. Next, we combined all 10,286 P-values corresponding to one genetic variant i into one overall P-value P_i using the Simes procedure [Simes, 1986], yielding 7,545,443 P-values P_i, one for each autosomal genetic variant tested.

This overall P-value per SNP indicates if an autosomal SNP influences DNA methylation *anywhere* on the X-chromosome, reducing this analysis to a GWAS for X-chromosomal DNA methylation. SNPs with an overall P-value $< 5 \times 10^{-8}$ were deemed significantly associated with X-chromosomal DNA methylation.

To identify independent effects among the identified variants, we performed iterative conditional analyses. We re-ran the entire above procedure, correcting for the strongest associated sentinel variant, as determined by the lowest overall P-value.

$$y_j = \beta_{ij}x_i + \gamma M + \delta U + \beta_{topSNP}x_{sentinel} \tag{4.2}$$

Having identified a new top SNP at the same genome-wide significance level of $P < 5x10^{-8}$, we again re-did our analysis, now correcting for two top SNPs. We repeated this process until no new independent effects were identified, which was after 47 such iterations, thus yielding 48 sentinel variants, corresponding to 48 different loci.

Next, to establish the female-specificity of the identified loci on X-chromosomal methylation, we aimed to validate the 48 identified loci in 1,398 males from the discovery cohorts for which the same genotype and methylation data were available. Any locus also having an effect in males would then mean that particular locus was not female specific. To do this, we tested the sentinel variant per locus found in females in the exact same way as we did in females, but also testing all SNPs within 1 Mb correlated to the sentinel variant ($R^2 \geq 0.8$ in males). A locus with any SNP having an overall P-value < 0.05 in males was not considered to be female-specific, yielding four loci with four corresponding sentinel variants.

Replication of sentinel variants associated with female-specific X-chromosomal DNA methylation in the replication cohorts

To replicate the four identified sentinel variants, we used an independent sample of 3,351 females from 5 different replication cohorts (see section Description of replication cohorts), all having genotype and 450k methylation data available. Similar to the discovery phase, each of four sentinel variants x_i was associated with all X-chromosomal CpGs y_j in each replication cohort k:

$$y_{jk} = \beta_{ijk}x_{ik} \tag{4.3}$$

each yielding a test-statistic t_{ijk}. We then combined the test-statistics corresponding to each genetic variant i and CpG j between each cohort k using Stouffer's weighted Z-method (discussed in Liptak [1958]), resulting in one overall Z-score Z_{ij} for each variant-CpG pair i, j:

$$Z_{ij} = \frac{\sum_k w_k t_{ijk}}{\sqrt{\sum_k w_k^2}} \tag{4.4}$$

where w_k indicates the sample size for replication cohort k. Converting each overall Z-score Z_{ij} to a P-value P_{ij}, we again used the Simes' procedure[Simes,

1986] to calculate one overall P-value P_i per genetic variant i, representing the statistical evidence for an association with any X-chromosomal CpG in the replication cohorts.

Local (*cis*) expression QTL mapping

In order to map the identified sentinel variants associated with female-specific X-chromosomal methylation to nearby genes, we employed *cis*-eQTL mapping in the discovery cohorts, where we associated the genotypes of a genetic variant with the expression levels of genes j *in cis* (< 250Kb). Similar to the *trans*-meQTL mapping for chromosome X, we corrected for known covariates M (*i.e.*, cell counts, cohort, age, technical batches), and unknown confounding U using *cate*, using an optimal number of latent factors to include, as suggested by *cate*:

$$y_j = \beta_{ij} x_i + \gamma M + \delta U \tag{4.5}$$

Next, we performed the Bonferroni correction to the corresponding P-values p_{ij} to identify genes associated with the genetic variant.

Associating X-chromosomal CpGs with changes in the expression of nearby genes

To identify genes associated with DNA methylation of nearby CpGs (< 250Kb), we used a similar model as for *trans*-meQTL and *cis*-eQTL mapping. We associated methylation levels of CpGs x_i with the observed expression values of a gene y_j using a linear model, correcting for covariates M (*i.e.*, cell counts, cohort, age, technical batches):

$$y_j = \beta_{ij} x_i + \gamma M \tag{4.6}$$

The Bonferroni correction was used to determine significant CpG-gene pairs.

Identifying genetic variants influencing autosomal DNA methylation

To identify long-range effects (> 5Mb) of a genetic variant on DNA methylation at autosomal CpGs, we performed *trans*-meQTL mapping using all 3,265 samples for which both genotype data and methylation data were available, as we expected these effects to be present in both women and men (Supplementary Fig. 6). For any genetic variant i and CpG j, we fitted a linear model correcting for known covariates M (cell counts, cohort, age, technical batches), and unknown confounding U using *cate*:

$$y_j = \beta_{ij} x_i + \gamma M + \delta U \tag{4.7}$$

The FDR was controlled within each set of corresponding P-values p_{ij}, to obtain a list of associated CpGs for a genetic variant i.

Testing epistatic effects

To test if the identified autosomal loci have any epistatic effects on X-chromosomal DNA methylation, we corrected the analysis for X-chromosomal *cis*-meQTLs. We first mapped *cis*-meQTLs (< 250Kb) on the X-chromosome by testing all nearby genetic variants for an effect on any of the X-chromosomal CpGs associated with one of the three autosomal loci. For any genetic variant i and CpG j, we fitted a linear model correcting for known covariates M (cell counts, cohort, age, technical batches):

$$y_j = \beta_{ij}^X x_i + \gamma M \tag{4.8}$$

We corrected for multiple testing using the Bonferroni procedure, selecting CpGs harboring *cis*-meQTLs. Next, we re-tested the effects of the autosomal loci on the X-chromosomal CpGs, but this time correcting for the strongest cis-SNP.

$$y_j = \beta_{ij}^X x_i^X + \beta_{ij}^{auto} x_i^{auto} + \gamma M \tag{4.9}$$

Annotations and enrichment tests

CpGs were annotation using UCSC Genome Browser [Kent et al., 2002], histone marks and chromatin states data from the Blueprint Epigenome data [Martens and Stunnenberg, 2013], transcription factor binding site (TFBS) data from the Encode Project [Consortium et al., 2012], and data on regions escaping X-inactivation [Cotton et al., 2014]. All annotations were done based on the location of the CpG site using HG19/GRCh37.

The CpG island (CGI) track from the UCSC Genome Browser was used to map CpGs to CGIs. Shores were defined as the flanking 2 kb regions. All other regions were defined as non-CGI.

We obtained Epigenomics Roadmap ChIP-seq data on histone marks measured in blood-related cell types (the GM12878 lymphoblastoid cell line, the K562 leukemia cell line, and monocytes). We selected five different histone marks for which data measured in both men and women were available (H3K4me3, H3K4me1, H3K9me3, H3K27me3, H3K27ac). A CpG was said to overlap with any histone mark if it did so in any the of the data sets.

We obtained Epigenomics Roadmap data on the 16 predicted core chromatin states data in blood-related cell types (the GM12878 lymphoblastoid cell line, the K562 leukemia cell line, and monocytes). A CpG was said to overlap with any chromatin state if it did so in any of the available data sets for that histone mark. Likewise, we obtained transcription factor binding data from the Encode Project, using blood-related cell types only (GM08714, GM10847, GM12878, GM12892, GM18505, GM18526, GM18951, GM19099, GM19193).

The degree of escape from X-inactivation for 632 transcription start sites (TSS) has previously been established in 27 different tissues [Cotton et al., 2014]. Within each tissue, each TSS was said to fully escape XCI, variably escape XCI, or be subject to XCI. We mapped each X-chromosomal CpG to the nearest such TSS, annotating each CpG with the accompanying scores for each of the 27

tissues. CpGs not in the vicinity of any such TSS (>10kb, 4,698 CpGs) were left unannotated.

In order to determine the enrichment of CpGs for any of the described genomic contexts, we used Fisher's exact test, where the used all X-chromosome CpGs as the background set.

Data availability

Raw data were submitted to the European Genome-phenome Archive (EGA) under accession EGAS00001001077 [https://www.ebi.ac.uk/ega/studies/EGAS00001001077].

References

Andrew, T. et al. [2001]. Are Twins and Singletons Comparable? A Study of Disease-related and Lifestyle Characteristics in Adult Women, *Twin Research* **4**(06): 464–477.

Aryee, M. J. et al. [2014]. Minfi: a flexible and comprehensive Bioconductor package for the analysis of Infinium DNA methylation microarrays, *Bioinformatics* **30**(10): 1363–1369.

Blewitt, M. E. et al. [2008]. SmcHD1, containing a structural-maintenance-of-chromosomes hinge domain, has a critical role in X inactivation, *Nature Genetics* **40**(5): 663–669.

Boomsma, D. I. et al. [2002]. Netherlands Twin Register: A Focus on Longitudinal Research, *Twin Research* **5**(5): 401–406.

Boyd, A. et al. [2013]. Cohort Profile: The 'Children of the 90s'—the index offspring of the Avon Longitudinal Study of Parents and Children, *International Journal of Epidemiology* **42**(1): 111–127.

Breiling, A. and Lyko, F. [2015]. Epigenetic regulatory functions of DNA modifications: 5-methylcytosine and beyond, *Epigenetics & Chromatin* **8**: 24.

Brinkman, A. B. et al. [2006]. Histone modification patterns associated with the human X chromosome, *EMBO Rep* **7**(6): 628–634.

Carrel, L. and Willard, H. F. [1999]. Heterogeneous gene expression from the inactive X chromosome: An X-linked gene that escapes X inactivation in some human cell lines but is inactivated in others, *Proceedings of the National Academy of Sciences* **96**(13): 7364–7369.

Carrel, L. and Willard, H. F. [2005]. X-inactivation profile reveals extensive variability in X-linked gene expression in females, *Nature* **434**(7031): 400–404.

Chen, K. et al. [2015]. Genome-wide binding and mechanistic analyses of Smchd1-mediated epigenetic regulation, *Proc Natl Acad Sci U S A* **112**(27): E3535–44.

Chen, Y. A. et al. [2013]. Discovery of cross-reactive probes and polymorphic CpGs in the Illumina Infinium HumanMethylation450 microarray, *Epigenetics* **8**(2): 203–209.

Chess, A. [2005]. Monoallelic expression of protocadherin genes., *Nature Genetics* **37**(2): 120–121.

Chu, C. et al. [2018]. Systematic Discovery of Xist RNA Binding Proteins, *Cell* **161**(2): 404–416.

Consortium, Dunham, I. et al. [2012]. An integrated encyclopedia of DNA elements in the human genome, *Nature* **489**(7414): 57–74.

Cotton, A. M. et al. [2013]. Analysis of expressed SNPs identifies variable extents of expression from the human inactive X chromosome, *Genome Biol* **14**(11): R122.

Cotton, A. M. et al. [2014]. Landscape of DNA methylation on the X chromosome reflects CpG density, functional chromatin state and X-chromosome inactivation, *Human Molecular Genetics* **24**(6): 1528–1539.

Datta, S. R. et al. [2008]. A threonine to isoleucine missense mutation in the pericentriolar material 1 gene is strongly associated with schizophrenia, *Molecular Psychiatry* **15**: 615.

Daxinger, L. et al. [2013]. An ENU mutagenesis screen identifies novel and known genes involved in epigenetic processes in the mouse, *Genome Biol* **14**(9): R96.

Deelen, J. et al. [2014a]. Genome-wide association meta-analysis of human longevity identifies a novel locus conferring survival beyond 90 years of age, *Human Molecular Genetics* **23**(16): 4420–4432.

Deelen, P. et al. [2014b]. Genotype harmonizer: automatic strand alignment and format conversion for genotype data integration, *BMC Research Notes* **7**(1): 901.

Dobin, A. et al. [2013]. STAR: ultrafast universal RNA-seq aligner, *Bioinformatics* **29**(1): 15–21.

Falckenhayn, C. et al. [2016]. Comprehensive DNA methylation analysis of the Aedes aegypti genome, *Scientific Reports* **6**: 36444.

Fortin, J. P. et al. [2014]. Functional normalization of 450k methylation array data improves replication in large cancer studies, *Genome Biol* **15**(12): 503.

Fraser, A. et al. [2013]. Cohort Profile: The Avon Longitudinal Study of Parents and Children: ALSPAC mothers cohort, *International Journal of Epidemiology* **42**(1): 97–110.

Galupa, R. and Heard, E. [2015]. X-chromosome inactivation: new insights into cis and trans regulation, *Current Opinion in Genetics & Development* **31**: 57–66.

Gaunt, T. R. et al. [2016]. Systematic identification of genetic influences on methylation across the human life course, *Genome Biol* **17**: 61.

Gendrel, A.-V. et al. [2012]. Smchd1-Dependent and -Independent Pathways Determine Developmental Dynamics of CpG Island Methylation on the Inactive X Chromosome, *Developmental Cell* **23**(2): 265–279.

Gendrel, A. et al. [2013]. Epigenetic Functions of Smchd1 Repress Gene Clusters on the Inactive X Chromosome and on Autosomes, *Molecular and Cellular Biology* **33**(16): 3150–3165.

Gilbert, W. V., Bell, T. A. and Schaening, C. [2016]. Messenger RNA modifications: Form, distribution, and function, *Science* **352**(6292): 1408 LP – 1412.

Hannon, E. et al. [2016]. An integrated genetic-epigenetic analysis of schizophrenia: evidence for co-localization of genetic associations and differential DNA methylation, *Genome Biology* **17**(1): 176.

Heard, E. et al. [2001]. Methylation of Histone H3 at Lys-9 Is an Early Mark on the X Chromosome during X Inactivation, *Cell* **107**(6): 727–738.

Hofman, A. et al. [2013]. The Rotterdam Study: 2014 objectives and design update, *European Journal of Epidemiology* **28**(11): 889–926.

Howie, B. N., Donnelly, P. and Marchini, J. [2009]. A Flexible and Accurate Genotype Imputation Method for the

Next Generation of Genome-Wide Association Studies, *plos genetics* **5**(6).

Huisman, M. H. et al. [2011]. Population based epidemiology of amyotrophic lateral sclerosis using capture-recapture methodology, *J Neurol Neurosurg Psychiatry* **82**(10): 1165–1170.

Ikram, M. A. et al. [2017]. The Rotterdam Study: 2018 update on objectives, design and main results, *European Journal of Epidemiology* **32**(9): 807–850.

Joshi Fass, J., N. [2011]. Sickle: a sliding-window, adaptive, quality-based trimming tool for FastQ files (version 1.33).

Kent, W. J. et al. [2002]. The Human Genome Browser at UCSC, *Genome Res* **12**(6): 996–1006.

Kundaje, A. et al. [2015]. Integrative analysis of 111 reference human epigenomes, *Nature* **518**(7539): 317–330.

Lemmers, R. J. et al. [2012]. Digenic inheritance of an SMCHD1 mutation and an FSHD-permissive D4Z4 allele causes facioscapulohumeral muscular dystrophy type 2, *Nature Genetics* **44**(12): 1370–1374.

Lin, B. D., Willemsen, G., Abdellaoui, A., Bartels, M., Ehli, E. A., Davies, G. E., Boomsma, D. I. and Hottenga, J. J. [2016]. The Genetic Overlap Between Hair and Eye Color, *Twin Research and Human Genetics* **19**(6): 595–599.

Liptak, T. [1958]. On the combination of independent tests, *Magyar Tud Akad Mat Kutato Int Kozl* **3**: 171–197.

Luijk, R. et al. [2015]. An alternative approach to multiple testing for methylation QTL mapping reduces the proportion of falsely identified CpGs, *Bioinformatics* **31**(3): 340–345.

Lyon, M. F. [1961]. Gene Action in the X-chromosome of the Mouse (Mus musculus L.), *Nature* **190**(4773): 372–373.

Martens, J. H. A. and Stunnenberg, H. G. [2013]. BLUEPRINT: mapping human blood cell epigenomes, *Haematologica* **98**(10): 1487–1489.

Martin, M. [2011]. Cutadapt removes adapter sequences from high-throughput sequencing reads, *EMBnet.journal* **17**(1): 10.

Mason, A. G. et al. [2017]. SMCHD1 regulates a limited set of gene clusters on autosomal chromosomes, *Skeletal Muscle* **7**(1): 12.

Massah, S. et al. [2014]. Epigenetic characterization of the growth hormone gene identifies SmcHD1 as a regulator of autosomal gene clusters, *PLoS One* **9**(5): e97535.

Min, J. et al. [2017]. Meffil: efficient normalisation and analysis of very large DNA methylation samples, *Doi.Org* p. 125963.

Moayyeri, A. et al. [2013]. Cohort Profile: TwinsUK and Healthy Ageing Twin Study, *International Journal of Epidemiology* **42**(1): 76–85.

Mould, A. W. et al. [2013]. Smchd1 regulates a subset of autosomal genes subject to monoallelic expression in addition to being critical for X inactivation, *Epigenetics Chromatin* **6**(1): 19.

Nozawa, R. S. et al. [2013]. Human inactive X chromosome is compacted through a PRC2-independent SMCHD1-HBiX1 pathway, *Nat Struct Mol Biol* **20**(5): 566–573.

Orru, V. et al. [2013]. Genetic variants regulating immune cell levels in health and disease, *Cell* **155**(1): 242–256.

Peeters, S. B., Cotton, A. M. and Brown, C. J. [2014]. Variable escape from X-chromosome inactivation: identifying factors that tip the scales towards expression, *Bioessays* **36**(8): 746–756.

Plath, K. et al. [2003]. Role of histone H3 lysine 27 methylation in X inactivation, *Science* **300**(5616): 131–135.

Roederer, M. et al. [2015]. The genetic architecture of the human immune system: a bioresource for autoimmunity and disease pathogenesis, *Cell* **161**(2): 387–403.

Sato, T., Okumura, F., Ariga, T. and Hatakeyama, S. [2012]. TRIM6 interacts with Myc and maintains the pluripotency of mouse embryonic stem cells, *J Cell Sci* **125**(Pt 6): 1544–1555.

Schoenmaker, M. et al. [2005]. Evidence of genetic enrichment for exceptional survival using a family approach: the Leiden Longevity Study, *European Journal of Human Genetics* .

Sharp, A. J., Stathaki, E., Migliavacca, E., Brahmachary, M., Montgomery, S. B., Dupre, Y. and Antonarakis, S. E. [2011]. DNA methylation profiles of human active and inactive X chromosomes, *Genome Res* **21**(10): 1592–1600.

Simes, R. J. [1986]. An Improved Bonferroni Procedure for Multiple Tests of Significance, *Biometrika* **73**(3): 751–754.

The Genome of the Netherlands Consortium et al. [2014]. Whole-genome sequence variation, population structure and demographic history of the Dutch population, *Nature Genetics* **46**(8): 818–825.

The International Schizophrenia Consortium [2008]. Rare chromosomal deletions and duplications increase risk of schizophrenia, *Nature* **455**(7210): 237–241.

Tigchelaar, E. F. et al. [2015]. Cohort profile: LifeLines DEEP, a prospective, general population cohort study in the northern Netherlands: study design and baseline characteristics, *BMJ Open* **5**(8): e006772.

Tobi, E. W., Slieker, R. C., Stein, A. D., Suchiman, H. E., Slagboom, P. E., van Zwet, E. W., Heijmans, B. T. and Lumey, L. H. [2015]. Early gestation as the critical time-window for changes in the prenatal environment to affect the adult human blood methylome, *Int J Epidemiol* **44**(4): 1211–1223.

van Greevenbroek, M. M. J. et al. [2011]. The cross-sectional association between insulin resistance and circulating complement C3 is partly explained by plasma alanine aminotransferase, independent of central obesity and

general inflammation (the CODAM study), *European Journal of Clinical Investigation* **41**(4): 372–379.

van Iterson, M. et al. [2014]. MethylAid: visual and interactive quality control of large Illumina 450k datasets, *Bioinformatics* **30**(23): 3435–3437.

Vaquerizas, J. M., Kummerfeld, S. K., Teichmann, S. A. and Luscombe, N. M. [2009]. A census of human transcription factors: function, expression and evolution, *Nat Rev Genet* **10**(4): 252–263.

Wang, J., Zhao, Q., Hastie, T. and Owen, A. B. [2015]. Confounder Adjustment in Multiple Hypothesis Testing, *ArXiv e-prints* .

Wu, T. P. et al. [2016]. DNA methylation on N6-adenine in mammalian embryonic stem cells, *Nature* **532**(7599): 329–333.

Yang, F., Babak, T., Shendure, J. and Disteche, C. M. [2010]. Global survey of escape from X inactivation by RNA-sequencing in mouse, *Genome Res* **20**(5): 614–622.

Yasukochi, Y. et al. [2010]. X chromosome-wide analyses of genomic DNA methylation states and gene expression in male and female neutrophils, *Proc Natl Acad Sci U S A* **107**(8): 3704–3709.

Zhang, Y. et al. [2013]. Genes that escape X-inactivation in humans have high intraspecific variability in expression, are associated with mental impairment but are not slow evolving, *Mol Biol Evol* **30**(12): 2588–2601.

Zhernakova, D. V. et al. [2017]. Identification of context-dependent expression quantitative trait loci in whole blood, *Nature Genetics* **49**(1): 139–145.