Dissertation
submitted to the
Combined Faculty of Natural Sciences and Mathematics
of the Ruperto Carola University Heidelberg, Germany
for the degree of
Doctor of Natural Sciences

presented by
M.Sc. Nan Li
born in: Heilongjiang, China
Oral examination: 8th December 2020

On the X chromosome inactivation of human induced pluripotent stem cells

Referees: Prof. Dr. Robert Russell
Prof. Dr. Benedikt Brors

Abstract

The human induced pluripotent stem cells (h-iPSCs) are valuable and promising tools for regenerative medicine and disease modelling because of their capacity to differentiate into multiple types of cells. For the application of female h-iPSCs, an important open question is whether they possess abnormal X chromosome inactivation (XCI) levels which might result in the alteration of gene expression and further downstream consequences.

This thesis investigates a population-level set of 273 female h-iPSCs from the Human Induced Pluripotent Stem Cell Initiative (HipSci) and shows a clear line-to-line variety in XCI levels, with four lines (1%) showing complete XCI loss. XCI level is associated with the expression of 2,086 genes (q-value < 0.1), 85% of which are on autosomes. XCI level is inherited in cells differentiated from h-iPSCs. Therefore, the variance of XCI might have an impact on downstream phenotypes, such as immune response. To allow researchers to quality control their h-iPSCs and to maximize the utility of existing h-iPSC banks, methylation-based and expression-based XCI metrics are proposed. These XCI metrics show a clear association between each other and can be used as covariates in further analysis.

To explore potential causal factors of XCI loss, variance component analyses are carried out with multiple potential sources, including donor information and technical or biological explanatory variables. These analyses reveal that culture time explains little of the XCI variation, that the expression of XIST is one of the most important explanatory factors, while still not a perfect marker, and that there is a significant donor effect.

To identify potential genetic determinants of XCI level, a genome-wide association study (GWAS) and a linear analysis with a subset of expression-related genetic variants are carried out. With cross-check of two XCI metrics, a variant region, as well as a single variant rs3790598, which is associated with the putative RNA helicase MOV10, are found as promising genetic sources of XCI variation.

Zusammenfassung

Die von Menschen induzierten pluripotenten Stammzellen (human indu-ced pluripotent stem cells, h-iPSCs) sind aufgrund ihrer Fähigkeit, sich in verschiedene Zelltypen zu differenzieren, wertvolle und vielversprechende Werkzeuge für die regenerative Medizin und die Modellierung von Krank-heiten. Für die Anwendung weiblicher h-iPSCs ist eine wichtige offene Frage, ob sie eine abnorme X-Chromosomen-Inaktivierung (XCI) aufweisen, die zu einer Veränderung der Genexpression und sogar zu weiteren nach-geschalteten Konsequenzen führen könnte.
Diese Dissertation untersucht 273 weibliche h-iPSCs von HipSci, einem Datensatz auf Bevölkerungsebene, und zeigt einen klaren Unterschied von Linien zu Linien bei den XCI-Niveaus, wobei vier Linien (1%) einen vollständigen XCI-Verlust aufweisen. Das XCI-Niveau ist mit der Expression von 2.086 Genen (q-Wert $< 0,1$) assoziiert, von denen 85% auf Autosomen liegen. Der XCI-Spiegel wird in Zellen vererbt, die sich von h-iPSCs differenzieren. Daher könnte die Variation von XCI einen Einfluss auf nachgelagerte Phänotypen, wie z.B. die Immunantwort, haben. Um Forschern die Qualitätskontrolle ihrer h-iPSCs zu ermöglichen und den Nutzen der bestehenden h-iPSC-Banken zu maximieren, werden methylierungs- und expressionsbasierte XCI-Metriken vorgeschlagen. Diese XCI-Metriken zeigen eine klare Assoziation untereinander und können als Kovariate in der weiteren Analyse verwendet werden.
Zur Untersuchung potenzieller kausaler Faktoren des XCI-Verlusts werden Varianzkomponentenanalysen mit mehreren potenziellen Ursachen durch-geführt, wozu Informationen zu Spendern und technische oder biologische Erklärungsvariablen gehören. Diese Analysen zeigen, dass die Zeit in der Kultur wenig von der XCI-Variation erklärt, dass die Expression von XIST einer der wichtigsten Faktoren zur Erklärung der XCI ist, obwohl sie noch immer kein perfekter Marker ist, und dass es einen signifikanten Spendereffekt gibt. Zur Identifizierung potenzieller genetischer Determinanten des XCI-Spiegels werden eine genomweite Assoziationsstudie (GWAS) und eine lineare Analyse mit einer Untergruppe von expressionsbezogenen genetischen Varianten durchgeführt. Bei der Gegenprobe von zwei XCI-Metriken auf Chromosom 1 werden eine Variantenregion sowie eine einzige Variante rs3790598, die mit der mutmaßlichen RNA-Helikase MOV10 assoziiert ist, als vielversprechende genetische Ursachen der XCI-Variation gefunden.

Acknowledgement

First and foremost, I would like to express my sincere thanks to my supervisor, Angela Teresa Filimon Gonçalves, for giving me the opportunity to work on this exciting project, as well as for all her supervision and support during my PhD. Being open-minded, responsible and passionate, she is a role model for my career. Warm thanks to the whole Gonçalves group, both past and present members. Especially to Roman Schefzik, for his guidance, support and discussions about statistical topics during my PhD. For the analysis of this PhD project, I would like to thank Daniel Gaffney for offering me the data access to Wellcome Trust Sanger Institute.

My Thesis Advisory Committee – Robert Russell, Benedikt Brors, Annette Kopp-Schneider and Wolfgang Huber, provided perceptive suggestions on my project, which I really appreciate. Special thanks to Benedikt Brors, who took the responsibility to supervise me during the time that my first supervisor, Bernd Fischer, passed away. I have been warmly treated by the whole Brors group and had intense discussions and meetings with scientists there, specially with Roman Kurilov, Abdelrahman Mahmoud and Jing Xu.

I would like to express my deep sorrow and insightful memories for Bernd Fischer, who was my first supervisor and passed away at a very young age. I would like to thank all friends, colleagues and the graduate school of DKFZ who supported me to go through that hard time.

Special thanks to my partner, Florian Sprenger, for giving me supports, understandings and encouragement all along my PhD during happy and hard times, as well as for his patient proofreading of this thesis.

Last but not least, I would like to thank my parents, who give me all their love and supports and always believe in my ability and creativity. Even though they are 7,000 km away from Heidelberg, I always feel their love which gives me the power to face all challenges in life and in the scientific research.

Contents

Acronyms

aIS allelic bias expression inactivation score.

bbs Bardet-Biedl syndrome.

chr chromosome.

CNA copy number alteration.

CV cross validation.

DP read depth (VCF file).

EBV Epstein Bar Virus.

eQTL expression quantitative trait loci.

FD feeder dependent (culture medium).

FF feeder free (culture medium).

GWAS genome wide association analysis.

h-ES human embryonic stem.

h-iPSC human induced pluripotent stem cell.

HipSci Human Induced Pluripotent Stem Cells Initiative.

HPC high performance computing.

INFg Interferon-gamma.

kb kilobase.

LCL lymphoblastoid cell line.

LD linkage disequilibrium.

m-iPSC mouse induced pluripotent stem cell.

MAF minor allele frequency.

mIS methylation inactivation score.

nd neonatal diabetes.

PCA principle component analysis.

PEER probabilistic estimation of expression residuals (method).

rIS expression ratio inactivation score.

RSS residuals sum of squares.

Rtt Rett syndrome.

SL1344 Salmonella typhimurium.

SNP single-nucleotide polymorphism.

VCA variance component analysis.

VIF variance inflation factor.

XCI X chromosome inactivation.

Xi inactive X chromosome.

Xic X-inactivation center.

XIST X-inactive specific transcript.

Chapter 1

Introduction

The scientific exploration is to keep answering open questions...and to keep asking new ones.

In 2006 and 2007, Yamanaka's lab published the induced pluripotent stem cell (iPSC) technology which successfully reprogrammed mouse embryonic stem (ES) cells and adult human dermal fibroblasts to pluripotent stem cells by transduction of four defined transcription factors: Oct3/4, Sox2, c-Myc, and Klf4 (Takahashi and Yamanaka 2006, Takahashi, Tanabe, et al. 2007). Since then, a series of works repeated the stable generation of human iPSCs (h-iPSC) and mouse iPSCs (m-iPSCs) with Yamanaka's method (e.g. Hu et al. 2010a, S. P. Paşca et al. 2011, Pomp et al. 2011, Mekhoubad et al. 2012) and nominated experimental alterations which improved the generation of h-iPSCs (Esteban et al. 2010, Zhao et al. 2008).

Since its development, iPSC technology has been of great interest for regenerative medicine because of its promises to derive multiple types of cells and for its tremendous potential for personalized cell therapy.

The Yamanaka lab initially proved that the h-iPSC cells had the capacity to differentiate into three germ layers, namely the endoderm, the mesoderm and the ectoderm (Takahashi, Tanabe, et al. 2007), followed by other groups which proved the differentiation ability of h-iPSCs to other cell types like neurons (Hu et al. 2010a, Schwartzentruber et al. 2018), macrophages (Alasoo et al. 2018, H. Zhang et al. 2015, Takahashi, Tanabe, et al. 2007), blood cells (Choi et al. 2009), or brain oligodendrocyte progenitor cells (S. Wang et al. 2013).

H-iPSC derived cells have been widely used in disease modelling, including the study of disease-related cellular phenotypes (S. P. Paşca et al. 2011, Y.-T. Lin et al. 2018), cell-malfunctions (Imaizumi et al. 2012), the regulation of antigen-receptors in tumor treatment (Y. Li et al. 2018), as well as the intervention for spinal cord injury (Tsuji et al. 2019).

For the application of h-iPSCs in biological and clinical research, open questions are whether there are genetic and/or expression variations in h-iPSCs from their original cells and whether there are effects of these variations (Martins-Taylor and R.-H. Xu 2012, Bilic et al. 2012). Previous studies discovered that h-iPSCs contain sub-chromosomal copy number variations (CNVs) (Chin et al. 2009, Spits et al. 2008), trisomy of chromosome 12 and chromosome X (Martins-Taylor, Nisler, et al. 2011, Taapken et al. 2011), as well as mutations relative to protein coding (Gore et al. 2011).

An important genetic feature for female h-iPSCs is X chromosome inactivation (XCI) status. XCI is the dosage compensation process in females that balances sex-related gene expression between males and females (details in section 1.2). Scientists have different observations and assumptions regarding the XCI in female h-iPSCs: some scientists observed two active X chromosomes and assumed that there is a X-reactivation during the programming of h-iPSCs (Barakat et al. 2015, Kim, Hysolli, Tanaka, et al. 2014); while other scientists reported a variable XCI level and assumed that a loss of XCI may happen on the inactive X chromosome in h-iPSCs during cell culture (Mekhoubad et al. 2012, Anguera et al. 2012, Tchieu et al. 2010, Brenes et al. 2020, Nazor et al. 2012).

Previous studies of XCI in h-iPSCs have limitations in following aspects: the limited number of h-iPSCs and donors for h-iPSCs used in the research (6 h-iPSCs generated from 2 different types fibroblasts in Kim, Hysolli, Tanaka, et al. 2014, 12 h-iPSCs in Mekhoubad et al. 2012, 30 h-iPSCs in Tchieu et al. 2010 and 11 donors in Trokovic et al. 2015), the generation of multiple h-iPSCs from the same donor (12 h-iPSCs from 2 patients in Pomp et al. 2011 and 7 h-iPSCs from 1 fibroblast line in Anguera et al. 2012), as well as multiple sources of h-iPSCs (69 h-iPSCs in Nazor et al. 2012 were generated from 7 institutes and from 3 reprogramming methods, namely 'Episomal', 'Lenti-virus' and 'Retro-virus').

This thesis investigates the XCI status in the data set of 273 female h-iPSCs from 205 independent donors generated by the Human Induced Pluripotent Stem Cell Initiative (HipSci, Kilpinen et al. 2017), including the prevalence of XCI level in the population, effects of experimental variables, potential genetic causes, as well as broad consequences of XCI variation in h-iPSCs and h-iPSC derived cells (Alasoo et al. 2018, Schwartzentruber et al. 2018).

HipSci is one of the largest data bank of h-iPSCs in the world generated by the same institute. There are three major advantages of using HipSci as the main data source of this thesis: the population-level recruited donors, the large-scale sample size and the uniform experimental and processing design. This uniform data source can help to reduce the noise and bias in h-iPSCs caused by different experimental methods (Newman et al. 2010, Volpato et al. 2018, Rao et al. 2012). HipSci recruits both healthy donors and patients from particular rare disease communities in the UK (www.hipsci.org). Compared with previous studies, using HipSci helps the understanding of XCI at population level. Moreover, for

68 out of 205 female donors, there is a second independently generated h-iPSC line available, making it possible to investigate the donor effect in the 'sibling' h-iPSC lines.

Besides HipSci, a smaller data set by Banovich et al. 2018 which contains 32 female h-iPSCs generated from YRI lymphoblastoid cell lines (African population in the 1000 Genome Project, 1000 Genomes Project Consortium et al. 2015) is used to reproduce the XCI-prevalence and the computation of XCI metrics, which demonstrates the existence of XCI heterogeneity in h-iPSCs regardless of the cell's origin.

1.1 HipSci, the world biggest h-iPSC bank from a single institute

The HipSci project was established by four key partners: the Wellcome Trust Sanger Institute, the European Bioinformatics Institute (EMBL-EBI), the King's college London and the University of Dundee. The motivation of the establishment of HipSci was to generate a large-scale, high quality h-iPSC reference base which is open to both academia and industry.
HipSci is a very important source for the study of h-iPSCs: firstly, mainly healthy donors were included in the project cohort, plus several donors with inherited genetic diseases, meanwhile the age and health condition (either healthy or with a certain genetic disease) were recorded; secondly, h-iPSC lines were generated from donors using a standardized experimental pipeline and went through quality control on their pluripotency; thirdly, each h-iPSC which passed the quality control was massively characterised on their genetics and genomics level, including whole exome sequencing, methylation array, RNA-sequencing, expression array and proteomics mass spectrometry (Kilpinen et al. 2017). Using h-iPSC lines from HipSci guarantees a uniform data source for the research. This is important since using h-iPSCs from different laboratories might introduce bias because of experimental settings and batch variables (Newman et al. 2010, Liang et al. 2013).

The major work of this thesis uses 273 h-iPSCs generated from 205 female donors for the investigation of XCI level and uses 219 male h-iPSCs as reference (figure 1.2). Since the HipSci project uses the same experimental pipeline, genetics and genomics screening methods, as well as data processing procedures, I summarize these technical details from Kilpinen et al. 2017 and present them in section 1.1.1.

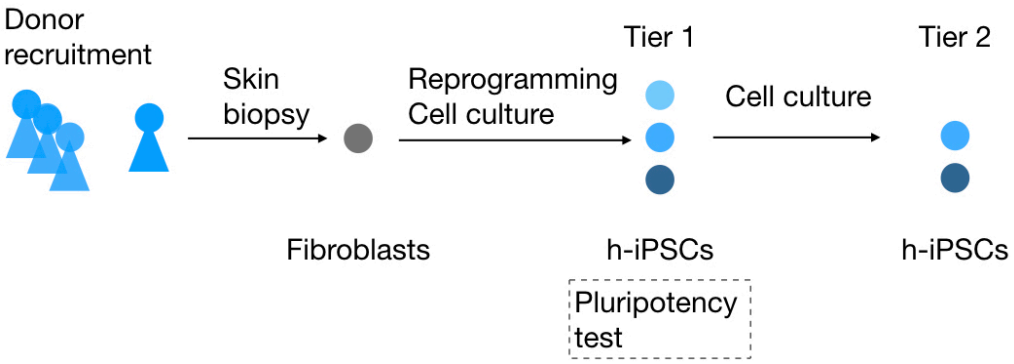


Figure 1.1: Summary of the generation process of h-iPSCs in HipSci (adapted from figure 1.a in Kilpinen et al. 2017).

1.1.1 Detailed experimental settings related to this thesis

Here, I summarize the 'Methods' and 'Supplementary Information' of Kilpinen et al. 2017 and present the information which is related to the analytical work of this thesis.

Generation and quality control of h-iPSCs

Volunteered donors of the HipSci project were recruited by the NIHR Cambridge BioResource and fibroblasts of each donor were obtained by skin punch biopsies. Yamanaka's method (Takahashi and Yamanaka 2006, Takahashi, Tanabe, et al. 2007) was applied for the reprogramming of h-iPSCs from fibrobalsts with transduction of human OCT3/4, SOX2, KLF4 and MYC using sendai vectors.

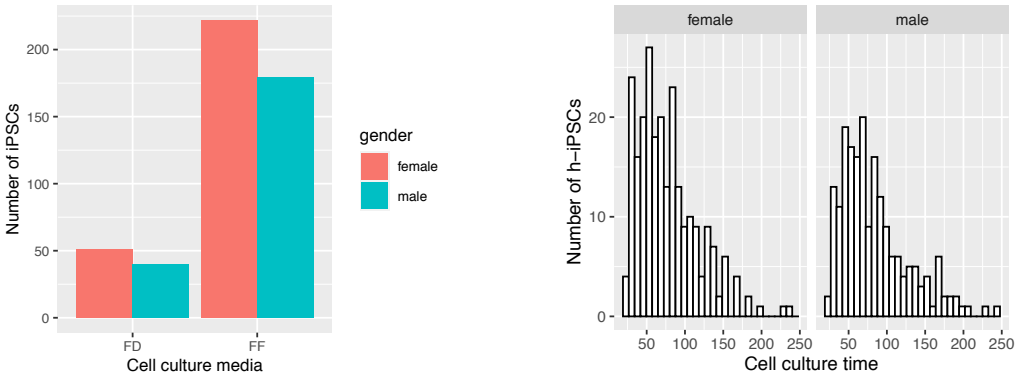
Quality control was executed with h-iPSCs' initial molecular data when they were passaged on average 16 times (Tier 1, figure 1.1). In details, criteria for the selection were: the level of pluripotency using the PluriTest assay (Müller et al. 2011), number of copy number abnormalities and ability to differentiate into each of three germ layers (endoderm, mesoderm and ectoderm). In the process of quality control, one or two lines were selected from the same donor to minimize the genetic abnormality between h-iPSCs and their progenitor fibroblasts.

This thesis includes 170 healthy donors, 21 donors with Bardet-Biedl syndrom (bbs) and 14 donors with neonatal diabetes (nd), and makes use of 273 female h-iPSC lines in total.

Important experimental differences in the cell culture of h-iPSCs

Culture media and culture time are two important experimental variables. The summary of these two experimental variables for all 492 h-iPSCs involved in

Figure 1.2: Summary of volunteered donors in the HipSci project



(a) (b)

Figure 1.3: Summary of important cell culture experimental variables for 492 h-iPSCs involved in this thesis. a. Two cell culture media are used: feeder free (FF) and feeder dependent (FD). Both male and female h-iPSCs have more lines cultured in medium FF than in medium FD. b. The cell culture time varies from 24 days to 240 days.

this thesis is shown in figure 1.3. For 273 female h-iPSCs used in this thesis, two types of media were used: feeder-free (FF, 222 lines) and feeder-dependent (FD, 51 lines). Besides, the time of cell culture was also different: the h-iPSCs were cultured from 24 days to 240 days (average cell culture time: 75 days). The 219 male h-iPSCs were cultured in the similar condition. Meanwhile the distribution of cell culture time was similar between male h-iPSCs and female h-iPSCs: the average culture time was 78 days in female h-iPSCs and was 84 days in male h-iPSCs (figure 1.3 b).

Multi-omics characterisation of h-iPSCs

All h-iPSCs in the HipSci were extensively screened. These data greatly help our understanding of h-iPSCs and allow the large scale quantitative analysis with h-iPSCs. Here, I present the screening methods, as well as major quality control process after the screening, which are associated with the analysis of this thesis, including the DNA methylation array, the RNA-sequencing, the expression array, as well as the genotyping array.

DNA methylation array

Two types of array were used for the measurement of DNA methylation level: the Illumina Human Methylation 450K (164 out of 273 female h-iPSCs and 125 out of 219 male h-iPSCs) and the Illumina Human Methylation 850K (109 out of 273 female h-iPSCs and 94 out of 219 male h-iPSCs). For these two methylation arrays, batch effects come from the plate where the sample is located (below, sample plate) and the exact position of this sample on the plate (below, sentrix). In HipSci, 11 different sample plates and 63 different sentrix for either male or female h-iPSCs were used.

The IlluminaHumanMethylation450kanno.ilmn12.hg19 Bioconductor annotation package was used for the probe annotation (Hansen 2016). The stratified quantile normalisation of samples was applied with *preprocessQuantile* function with minfi Bioconductor package (Aryee et al. 2014). In total 9,257 probes are located on the X chromosome.

RNA-sequencing

The RNA-sequencing of h-iPSCs in the HipSci project was executed using Illumina HiSeq 2000 system (75-base paired-end). The alignment of raw RNA-sequencing reads was done using STAR, version 2.4.0 (Dobin et al. 2013), with human reference GRCh37 (ENSEMBL 2010). Mapped reads were quantified at gene-level using HTSeq, version 0.6.1p1 (Anders et al. 2015) and were annotated against Genecode version 19 (Harrow et al. 2012).

In HipSci, in total 54,410 genes were screened by RNA-sequencing, including

2,190 X-located genes, 286 Y-located genes and 51,934 autosomal genes. The RNA-sequencing data was available for all 492 h-iPSCs involved in this thesis.

Expression array

Gene expression profiles were measured by Illumina Human HT-12 v4 Expression BeadChips. Probes were re-mapped against the human genome GRCh37 (Harrow et al. 2012) using BWA, version 0.7.5 (Heng Li and Durbin 2009). Mapped probes were filtered at two levels: probes whose minimum filtering quality (MAPQ) was smaller than 10 and probes which overlapped with any variant with minor alle frequency (MAF) greater than 0.05 in the main imputed data set were removed. After filtering, 25,604 probes remained, standing for 17,116 unique genes. Variance stabilization of the expression array was carried out with R/Bioconductor package vsn (Huber et al. 2002).

Genotyping and copy number alteration (CNA)

The genotypes of h-iPSC lines and fibroblasts were measured by an Illumina HumanCoreExome-12 BeadChip. The internal Illumina Genome Studio software was used for genotype calling. Following the initial quality control in the Illumina system, the imputation of genotype was done with IMPUTE2, version 2.3.1 (Howie et al. 2009), then haplotype estimation was carried out with SHAPEIT, version 2.r790 (Delaneau et al. 2012). VCF files from single samples were merged together while the INFO score was recalculated with posterior probabilities of genotypes. Another quality control was executed so that variants with the INFO score smaller than 0.4 were excluded.

Copy number alterations (CNAs) between h-iPSCs and fibroblasts of the same donor were called using the cnv function of Bcftools, version 1.9 (Heng Li 2011, Danecek, S. A. McCarthy, et al. 2016). The filtering process of CNAs was executed at three levels: the quality score (not smaller than 2), the number of deletions of markers (not smaller than 10) and the number of duplications of heterozygous markers (not smaller than 10). To summarize the detected CNA level in the initial result (711 h-iPSCs), 18% of h-iPSCs contained one or more CNA, meanwhile, 22% of CNAs were observed in at least one line generated from the same donor and 15% were observed in all replicates.

1.2 XCI, an open topic for h-iPSCs

XCI is a dosage compensation mechanism in mammals that balances the sex-related genes in the two genders: since females have two copies of X chromosome (XX) while males have one copy of X chromosome and one copy of Y chromosome (XY), XCI ensures that both females and males have similar expression level of the X-located genes (Lyon 1961, Brockdorff et al. 2015, Heard et al. 1997,

Avner et al. 2001).

The XCI process controls the silencing of the X chromosome in females and is a key process in the development of early embryos. The initialization time of XCI in humans is still not clearly known as in mouse, in which XCI is initiated at the preimplantation stage following early whole-genome activation (Huynh et al. 2003, Okamoto et al. 2004, Erhardt et al. 2003, Costanzi et al. 2000, Hartshorn et al. 2003, Johnson et al. 2004, Zernicka-Goetz 2002).

In both humans and mouse, the XCI process is believed to be controlled via the X-chromosome inactivation center (Xic), which is found mandatory for this process (J. T. Lee and Jaenisch 1997, J. Lee et al. 1999). The non-coding RNA XIST, which is the abbreviation of the X inactive-specific transcript, is the most important component of Xic and is seen as the key factor for the start of the XCI process (Penny et al. 1996, C. J. Brown, Hendrich, et al. 1992, Berg et al. 2009, Avner et al. 2001).

1.2.1 Previous studies have various conclusions about XCI level in h-iPSCs

Most of our knowledge about XCI in iPSCs comes from mouse iPSC (m-iPSC) lines: during the generation of m-iPSCs, the inactive X chromosome in mouse stem cells get reactivated, thus two active X chromosomes can be observed in m-iPSCs; once the m-iPSCs are differentiated to other cell types, the XCI takes place to ensure that there is only one active X in the mouse stem cells (Lyon 1961, Van den Berg et al. 2011, Liang et al. 2013, J. T. Lee and Bartolomei 2013, Galupa et al. 2018, Janiszewski et al. 2019, Pasque et al. 2015).

However, the XCI regulation in human iPSCs (h-iPSCs) is still unclear. As presented at the beginning of this chapter, scientists have different observations in previous studies: some scientists observe two active X chromosomes and believe that in h-iPSCs there is also a reactivation of the X chromosome as in m-iPSCs (Kim, Hysolli, and Park 2011, Tomoda et al. 2012, Barakat et al. 2015, Vacca et al. 2016); while other scientists observe a loss of XCI level which might result from the cell culture (Kim, Hysolli, Tanaka, et al. 2014, Mekhoubad et al. 2012, Anguera et al. 2012, Pomp et al. 2011, Nazor et al. 2012).

Considering that h-iPSCs and iPSC-derived cells are widely used in disease modeling and personalized cell therapy (S. P. Paşca et al. 2011, Hao Wu et al. 2014, S. Wang et al. 2013, Brix et al. 2005), to control the bias in research and in clinical applications, it is essential to clarify the general status of XCI in h-iPSCs, its inheritance in iPSC-derived and its consequences.
For example, a second active X chromosome might lead to misregulation of gene expression, which is problematic in biological functions. Besides, since the X chromosome contains the largest number of immune-related genes in human

(Bianchi et al. 2012a, Libert et al. 2010), a second active X chromosome might also result in immunological consequences.

1.2.2 XIST, the critical factor of the XCI process

The long non-coding RNA (lncRNA) XIST is identified as the mandatory factor for the initialization of the XCI process in humans by researches to date (Pontier et al. 2011, Simon et al. 2013, Galupa et al. 2018). In human embroynic stem (h-ES) cells, XIST locates on both two X chromosomes and expresses only on the inactive X chromosome (Xi, C. J. Brown, Hendrich, et al. 1992). The transcript of XIST remains in the nucleus of the stem cell and coats the Xi (Avner et al. 2001, C. J. Brown, Hendrich, et al. 1992). After XIST coats the Xi, the repressive marks (e.g. histone H3 lysine 9 dimethylation) accumulate and silencing genes (Polycombcomplex 1 and 2) are recruited (Maduro et al. 2016, Galupa et al. 2018). Several studies have presented the important role of DNA methylation in maintaining the silence of Xi (Panning et al. 1996, Hellman et al. 2007, Tribioli et al. 1992). Furthermore, once established, Xi is maintained and stably inherited upon cell divisions (Duncan et al. 2018, Maduro et al. 2016, Galupa et al. 2018).

XIST is not the only key factor in the establishment of XCI. In fact, XIST and its surrounding neighbourhoods formulate the X-inactivation center (Xic), while the exact sequence of Xic in h-ES cells and its detailed mechanism is still unclear (J. T. Lee and Jaenisch 1997, J. Lee et al. 1999). What is also unclear in terms of XIST is how the XIST transcript binds to the Xi. So far, scientists believe that XIST-RNA is a structural RNA in nucleus and it associates loosely with the nuclear matrix (Clemson et al. 1996, C. J. Brown, Hendrich, et al. 1992).

The XIST expression level in m-iPSCs has been found associated with culture time (Janiszewski et al. 2019). The recent study by Briggs et al. 2015 with single-cell technology also shows the loss of XIST expression in h-iPSCs with cell culture. These observations reveal the possibility that the XIST expression level, or even the XCI status which is regulated by XIST, might change over time in cell culture.

With the biological knowledge and the unsolved problems about the h-iPSC generation and about the XCI, I have the following focus in the research design: the X-methylation level, which is the direct representation of XCI; the XIST expression level, which controls and initializes the XCI; the time-effect in X-methylation and in XIST-expression; as well as the X-located and autosomal gene expression level, which might be major consequences of the XCI variation.

1.2.3 XCI escapees: the X-genes which are able to escape the XCI

On the X chromosome, a fraction of genes are able to escape from the XCI, which are named XCI escapees. A typical XCI escapee is the XIST, as described in sec-

tion 1.2.2, XIST expresses on the inactive X chromosome (Xi), which is seen as the initialization of the XCI (C. J. Brown, Hendrich, et al. 1992,Pontier et al. 2011, Simon et al. 2013, Galupa et al. 2018).

Besides XIST, a specific region is known to be XCI-escapees: the pseudoautosomal regions (PAR1 and PAR2, Helena Mangs et al. 2007, Raudsepp et al. 2015). The pseudoautosomal regions locate on both termini of the X chromosome and the Y chromosome and recombine during the male meiosis (Helena Mangs et al. 2007, Balaton et al. 2015). Therefore, genes in the pseudoautosomal regions have identical expression level between males (XY) and females (XX) and do not need further dosage compensation (Helena Mangs et al. 2007, Raudsepp et al. 2015, Balaton et al. 2015).

In human females, up to 25% of genes are able to escape from XCI, identified by Carrel et al. 2005 with human fibroblasts: 15% of genes can escapes in all samples while 10% of genes have a variable escape-pattern across samples. Furthermore, the XCI escapees in human are found to be tissue specific by Tukiainen et al. 2017 which investigated 29 types of tissues in 449 individuals and Cotton et al. 2015 which studied 4 types of tissues in 95 individuals.

Since there is a lack of hard proof about which genes are h-iPSC specific XCI escapees, this thesis uses the list of genes which escaped XCI in all tissues ($n = 99$) in Tukiainen et al. 2017 as a 'strict' representation of XCI escapees (chapter 2), and the list of genes which escaped XCI in at least one tissue ($n = 200$) in Tukiainen et al. 2017 as a 'loose' representation of the XCI escapees (chapter 5). Genes in these two lists are checked in the analysis to remove the effect of known XCI escapees in the result.

1.3 The motivation and research design of this thesis

The first and the most important question that I want to answer with this thesis is whether there is a variable XCI status in h-iPSCs from healthy donors at population level. With the h-iPSCs generated and screened by HipSci, I use the methylation level of the X chromosome as the direct read-out of the XCI and firstly present its overview in all female h-iPSCs (chapter 2). To facilitate the estimation of XCI status, I show that the expression level can also be used as XCI metric (chapter 2). Potential sources of XCI variation, including experimental variables, donor age and health condition, XIST expression level as well as the copy number alterations (CNAs), are estimated for their contribution to the XCI variability (chapter 3). Autosomal genetic variants are also studied for their association with XCI level (chapter 4). Besides, I also investigate the consequences of XCI variation in h-iPSCs and its inheritance in iPSC-derived cells (chapter 5). Finally, using a smaller data set of h-iPSCs generated from lymphoblastoid cell lines (LCLs, $n = 32$), some part of analysis is reproduced to demonstrate the

general XCI heterogeneity in h-iPSCs (chapter 6).

In this thesis, I carry out analytical research to answer open questions about the XCI status in h-iPSCs. I expect that the work of this thesis will not only expand the current biological knowledge about the h-iPSCs but also help the clinical application of h-iPSCs and iPSC-derived cells.

Chapter 2

Landscape of XCI in female h-iPSCs

Is there variation of XCI level in female h-iPSCs?
Yes, at population level.

In this chapter, I answer the first and the most important question related to the work of this thesis: what variation of XCI level is observable with female h-iPSCs in HipSci, the population-level h-iPSC data bank?

As presented in chapter 1, regarding the XCI level in female h-iPSCs, previous studies had different conclusions: some studies showed stable XCI status (Tchieu et al. 2010, Pomp et al. 2011) while others showed a various XCI level (Kim, Hysolli, and Park 2011, Marchetto et al. 2010, Tomoda et al. 2012, Barakat et al. 2015, Mekhoubad et al. 2012, Nazor et al. 2012). There are three major limitations of previous studies referring to the prevalence of XCI in h-iPSCs: the limited number of samples (typically within 30 h-iPSC lines and 15 donors), generation of multiple h-iPSCs from the same donor, as well as the usage of h-iPSCs from multiple sources (chapter 1).

Thanks to HipSci (Kilpinen et al. 2017), this thesis is able to present the XCI level in h-iPSCs using 273 female h-iPSC lines from 205 independent donors, with the majority of lines (235 out of 273) generated from healthy donors. Meanwhile, all h-iPSCs were intensively screened for their multi-omics characters so that I am able to present the overview of XCI with the methylation level of the X chromosome, to summarize it with multiple screening results and to present the consistency of XCI metrics obtained from different screening data in this chapter.

Since methylation is part of the mechanism of the silencing of the X chromosome (Galupa et al. 2018), results from the methylation array (chapter 1) were firstly used for the overview of XCI level in 273 female h-iPSCs (section 2.1), where I show the existence of the line to line variability in XCI level for female h-iPSCs.

To have a direct representation of XCI for a certain h-iPSC instead of using the whole methylation matrix, section 2.2 introduces the definition of methylation inactivation score (mIS) as XCI metric. To facilitate the estimation of XCI level of h-iPSCs for laboratories where methylation level is not measured, I show that expression level can also be used as XCI metrics which have good association with mIS (section 2.3). Furthermore, using genes selected by penalized linear regression can not improve the estimation of XCI level in h-iPSCs (section 2.5). With the single cell data set generated by Linker et al. 2019, I present that single cells display similar XCI level to the bulk level for female h-iPSC line joxm_1 (section 2.4).

2.1 Line-to-line variability of methylation level in female h-iPSCs

As briefly introduced before, results from the methylation array were used for the overview of X-chromosome methylation level in h-iPSCs. In the HipSci project, the methylation level of all h-iPSCs were screened by either Illumina Infinium Human Methylation 450K BeadChip or Illumina Infinium Human Methylation 850K BeadChip, including 9,257 probes on the X chromosome.

2.1.1 Line-to-Line variability of methylation level in female h-iPSCs

To present the methylation level at each locus of the X chromosome, this section introduces β value which represents the fraction of DNA molecules methylated at a certain locus, with the formula:

$$\beta = \frac{M}{U + M + 100} \tag{2.1}$$

where M =methylated allele intensity, U = unmethylated allele intensity.

Usually, β falls into interval $[0, 1)$. For probes on the X chromosome, a value of $\beta = 0$ indicates that this locus is unmethylated in all molecules. When $\beta > 0$, a certain proportion of molecules is methylated at this locus, for instance, $\beta = 0.5$ indicates that half of molecules is methylated. While a single β value presents the fraction of methylation at a locus, for each h-iPSC, the distribution of β on all X-located loci indicates the pattern of methylation in this h-iPSC line.

By presenting the β on all loci of X chromosome in all studied female h-iPSCs and in a subset of male h-iPSCs, figure 2.1 clearly shows the line-to-line variability of the methylation level in the female h-iPSC population.

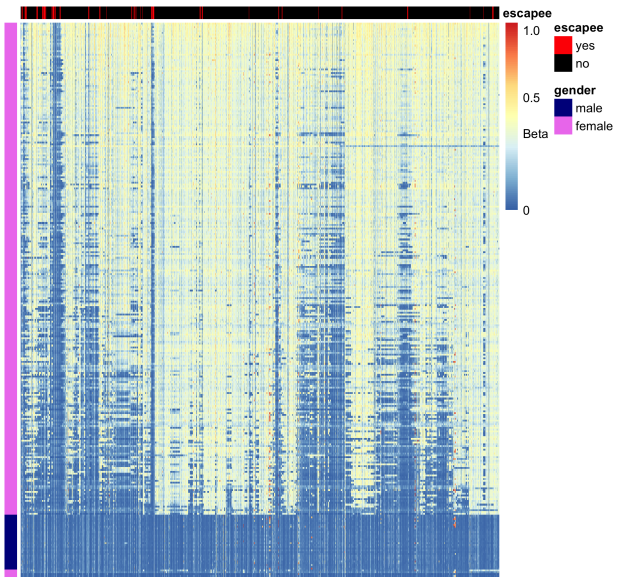


Figure 2.1: Heatmap of β on the X chromosome in 273 female h-iPSCs and 30 representative male h-iPSCs. H-iPSCs are placed in rows, labeled by their gender (blue: male, purple: female). The four outlier female h-iPSCs which display complete XCI loss are placed in the bottom rows. Probes are placed in columns, ordered by their genomic positions. Probes which located in regions of known XCI escapees are shown in red. β is written as Beta.

In figure 2.1, all 273 female h-iPSCs (in purple) and a representative subset of 30 male h-iPSCs (in blue) are placed in rows; probes on X chromosome and Y chromosome are placed in columns and ordered by their genomic position. A selection of probes were carried out for this figure: probes which were highly methylated (median methylation $\geqslant 0.4$) in both female and male h-iPSCs were removed since they were not relevant to the XCI (total number of probes in the figure: $4,285$). As briefly introduced in chapter 1 (section 1.2.3), genes which escaped XCI in all tissues in Tukiainen et al. 2017 ($n = 99$) are used as a 'strict' representation of XCI escapees. Probes which located in regions of these genes are labeled in red ($n = 235$). Among all escapee-targeting probes, 49% of probes display relatively high methylation level ($\beta > 0.25$) while 51% of probes display low methylation level ($\beta \leqslant 0.25$), indicating that in h-iPSCs, XCI might also happen on genes where were identified as escapees in other human tissues (Supplementary figure B.2).

Figure 2.1 clearly reveals the variance of β in female h-iPSCs: on one hand, for a certain locus, different female h-iPSCs have different β value; on the other hand, different female h-iPSCs have different overall β across the X chromosome. This observation demonstrates the existence of XCI variability in the population of female h-iPSCs from healthy donors by showing the variant methylation level of the X chromosome.

2.1.2 Three patterns of methylation level in female h-iPSCs

For all 273 female h-iPSCs studied, three patterns of distribution were observed for β on the X chromosome: only one peak at $\beta = 0.5$, indicating that for this h-iPSC half of X chromosome is methylated, which is the sign of proper XCI; only one peak at $\beta = 0$, indicating that all X chromosome is unmethylated, which is the sign of complete XCI loss; as well as two peaks at $\beta = 0$ and at $\beta = 0.5$, indicating that this h-iPSC has part of X chromosome methylated, which is the sign of incomplete XCI loss. $\beta = 1$ refers to the base of the methylation array so is not included for the discussion and the analysis. These three patterns of methylation level indicate the three different XCI level in the female h-iPSCs. One example of each pattern is shown in figure 2.2 (a, b and c). In contrast with the variable methylation level in female h-iPSCs, all 219 male h-iPSCs showed a uniform distribution of β: only one peak at $\beta = 0$, shown in figure 2.2 (d). To summarize, among the 273 female h-iPSCs, 4 lines (1%) have complete XCI loss while the majority shows variable XCI level.

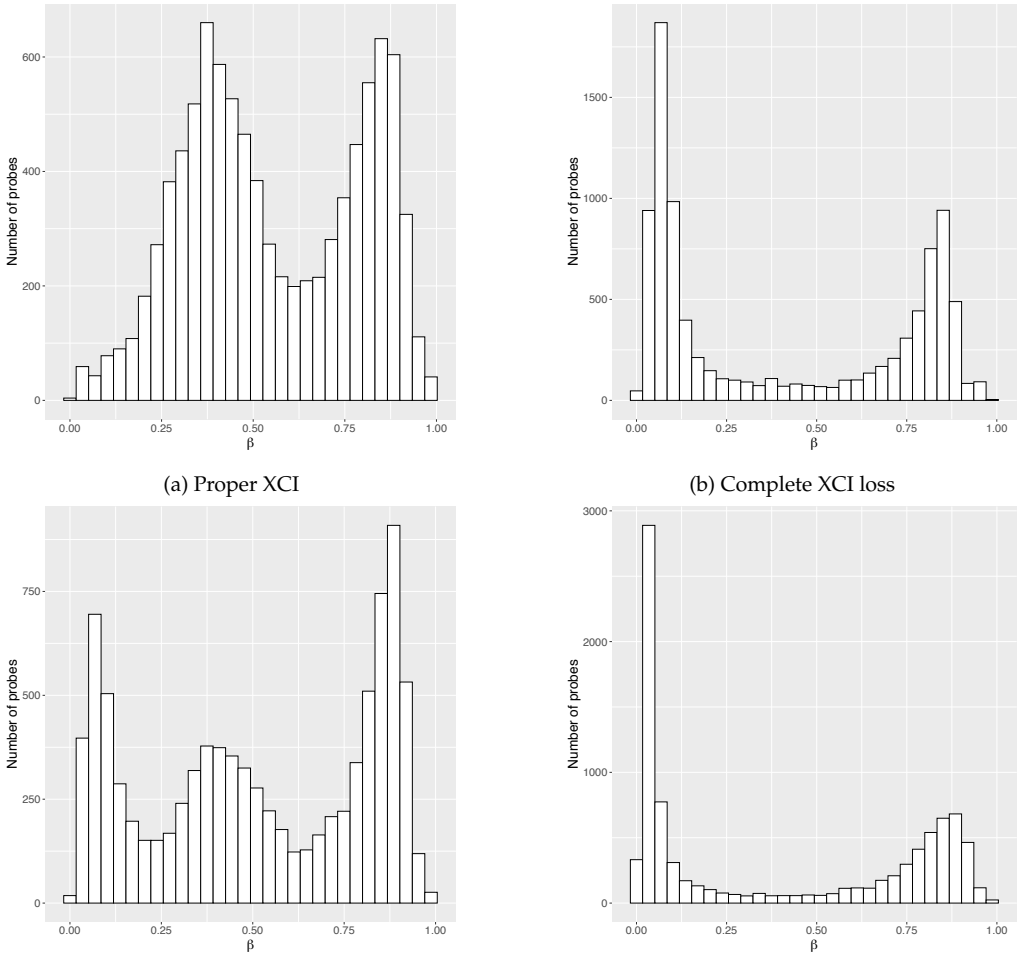
2.1.3 Lines from the same donor display a similar XCI level

Many previous studies used multiple h-iPSCs generated from the same donor (Mekhoubad et al. 2012, Anguera et al. 2012, Pomp et al. 2011), whereas the question arose: regarding the XCI level, are lines from the same donor (below, sibling lines) more similar to one another than lines from different donors?

In HipSci (Kilpinen et al. 2017), a second independent h-iPSC line is available for 68 out of 205 female donors. In the cell culture, it happened that 50 out of 68 donors had both two sibling-lines cultured with medium Feeder Free (FF), 8 donors had both two sibling-lines cultured with medium Feeder Dependent (FD), while 10 donors had sibling-lines cultured in different media. To control the experimental bias, I take 50 pairs of sibling lines which were generated in medium FF, as well as 129 h-iPSCs from independent donors generated in the same medium, then compute the correlation of β (formula 2.1) between paired sibling lines and among all independent h-iPSCs (figure 2.3).
Figure 2.3 shows that the methylation level in sibling lines is higher correlated (mean $= 0.94$, median $= 0.95$) than h-iPSCs from independent donors which were generated in the same experimental condition (mean $= 0.91$, median $= 0.91$). This observation demonstrates that sibling lines are more similar to each other when compared with h-iPSCs from different donors.

2.2 Methylation-based XCI metric: methylation inactivation score

The methylation inactivation score (mIS) is a summary of β values (formula 2.1) over the X chromosome. As methylation of the X chromosome is directly related to the XCI process, in this thesis, mIS is used as a standard representation of XCI.



(a) Proper XCI

(b) Complete XCI loss

(c) Incomplete XCI

(d) Male h-iPSC

Figure 2.2: Distribution of X-related β indicates different methylation patterns in female and male h-iPSCs. a, b and c. Three patterns of distributions of β in female h-iPSCs, representing the proper XCI, complete XCI loss and incomplete XCI loss (airc_66, dons_1 and fpdj_3, respectively). d. An example of the uniform distribution of β in male h-iPSCs (ffdc_1).

2.2.1 Definition of mIS

The mIS is defined as log transferred number of probes (loci) on the X chromosome which are unmethylated, with formula:

$$\text{mIS} \coloneqq \log_{10}(\text{number of loci on X chromosome with } \beta < 0.25) \tag{2.2}$$

The mIS directly measures, for a certain h-iPSC line, how many probes on X chromosome are unmethylated (threshold at $\beta = 0.25$). To identify whether $\beta = 0.25$ could separate 'unmethylated probes' and 'methylated probes', a k-means clustering, with k= 3 (represents for potential peak at 0, at 0.5 and at 1),

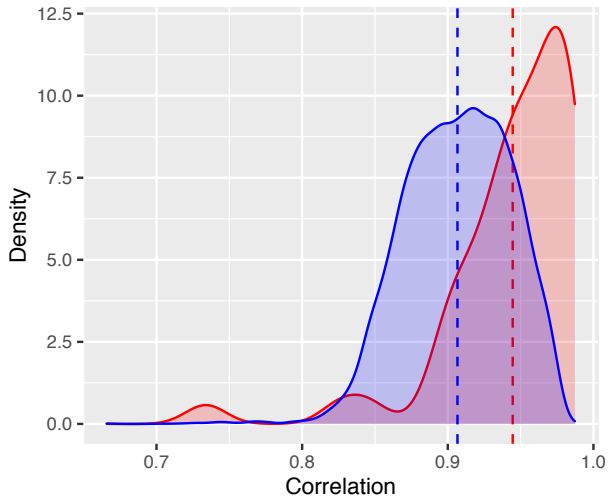


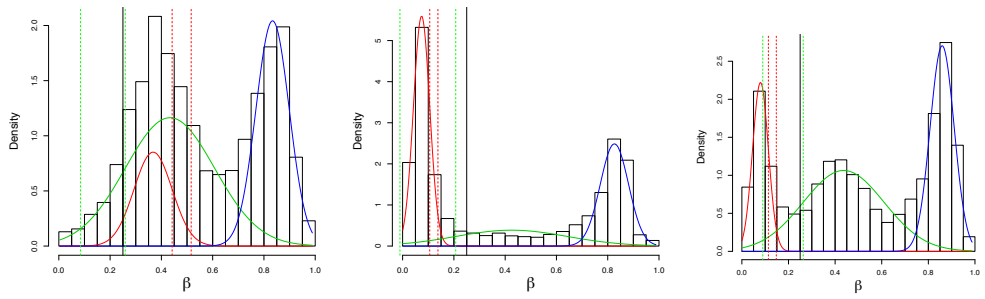
Figure 2.3: Comparison of the correlation of β values in 129 independent h-iPSCs (blue) and in 50 pairs of sibling h-iPSCs (red). Dashed lines label the average of the correlation: 0.91 (blue) and 0.94 (red).

was applied to X-related β values in female h-iPSCs. Using the same female h-iPSC lines in figure 2.2, figure 2.4 shows the distribution of β with the density curve of three clusters.

Figure 2.4 shows that, for female h-iPSCs with XCI loss pattern as complete or incomplete (figure 2.4 b and figure 2.4 c), $\beta = 0.25$ is able to separate two clusters which represent unmethylation (peak at $\beta = 0$) and methylation (peak at $\beta = 0.5$). For h-iPSCs with proper XCI (overlapping two clusters with peak at $\beta = 0.5$, figure 2.4 a), $\beta = 0.25$ separates a small proportion of probes which are unmethylated from major probes with proper methylation rate. The same distribution plot with k-means clustering was applied to all 273 female h-iPSCs, of which results showed a proper separation with $\beta = 0.25$. As the fixed β value is able to serve as a proper threshold, it is unnecessary to apply a line-specific threshold.

2.2.2 Correction of technical factors in methylation array

In section 1.1.1, technical details of DNA methylation array in HipSci (Kilpinen et al. 2017) was presented. To briefly summarize, two types of array were applied: Illumina Human Methylation 450K (164 out of 273 female h-iPSCs) and Illumina Human Methylation 850K (109 out of 273 female h-iPSCs). In the methylation array, two other technical factors were also different for female h-iPSC lines: the plate on which the sample is placed (below, sample plate, total number = 11) and the position of the sample on the plate (below, sentrix ID, total number = 61).



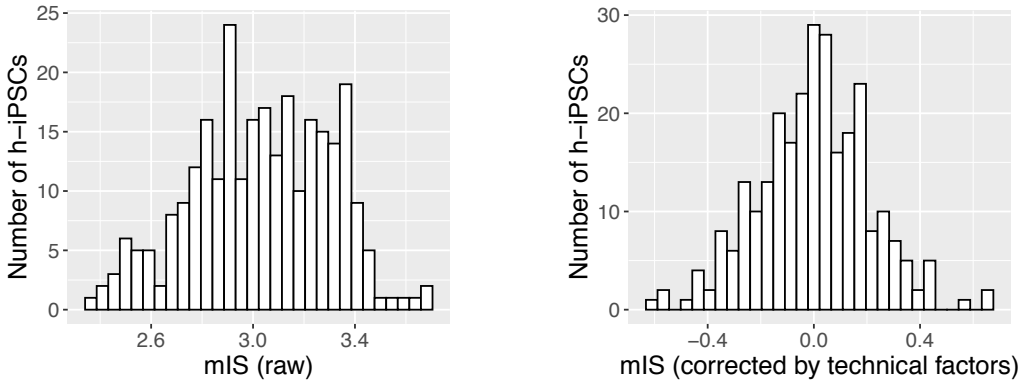
(a) Proper XCI (b) Complete XCI loss (c) Incomplete XCI

Figure 2.4: Histogram of β with density curves in 3-clusters (colors: red for unmethylation cluster, green for proper XCI cluster, blue for technology base; dash curves are density of clusters; vertical red and green dash lines are first and second standard deviation for the representing cluster; vertical black line is threshold $\beta = 0.25$). a. An example of proper XCI (airc_66). b. An example of complete XCI loss (dons_1). c. An example of incomplete XCI (fpdj_3).

To adjust these technical factors, a linear model is fit, with formula:

$$\text{mIS} = 0 + \text{sentrix ID} + \text{sample plate} + \text{array type}. \tag{2.3}$$

In formula 2.3, all three technical factors are categorical. Residuals of formula 2.3 are used as technical-corrected mIS values. The comparison of raw and technical-corrected mIS for 273 female h-iPSCs is shown in figure 2.5.



(a) (b)

Figure 2.5: The distribution of mIS in 273 female h-iPSCs before (a) and after (b) correction for three technical factors in DNA methylation array (type of array, ID of sentrix and sample plate).

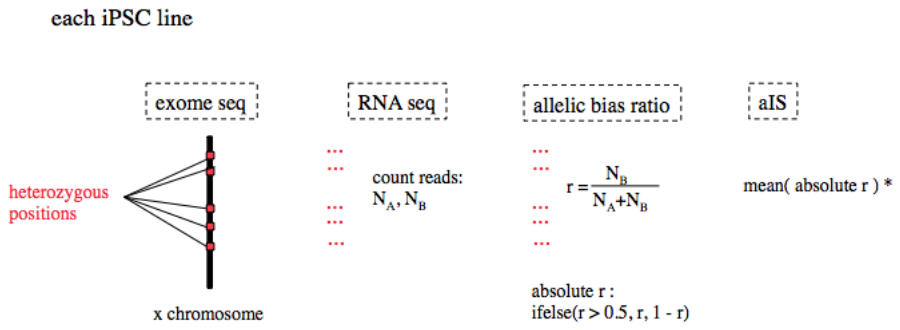


Figure 2.6: Computation of aIS

2.3 Expression based XCI metrics

It is common practice for laboratories working with h-iPSCs to measure the expression level as it is part of pluripotency assay (Müller et al. 2011). To facilitate the estimation of XCI status in female h-iPSCs, here I present two expression-based XCI metrics and show that they are well associated with the methylation-based XCI metric, mIS.

In this study, expression level from both RNA-sequencing and microarray assays are explored. Technical details and normalization process of these two screens were presented in section 1.1.1. The two expression-based XCI metrics are: the mean allelic bias expression inactivation score (aIS) and the expression ratio inactivation score (rIS). For the computation of aIS, the genotyping data and RNA-sequencing data are necessary; while for rIS, only expression data are needed.

2.3.1 Definition and computation of aIS

The aIS is defined as the average of ratio of alternative allele expression on heterozygous positions on the X chromosome (figure 2.6). This exploration is inspired by the assumption that following loss of XCI, a bi-allelic expression on heterozygous positions of the X chromosome can be observed.
As described in figure 2.6, for each h-iPSC line, the heterozygous positions on the X chromosome are obtained by exome-sequencing data as described in Kilpinen et al. 2017 (details in section 1.1.1). With RNA-sequencing data of one h-iPSC, for each heterozygous position on the X chromosome, numbers of reads of reference allele and alternative allele are counted (N_A and N_B respectively).

The allelic bias ratio, r, is defined as the ratio of alternative allele reads to the sum of reference and alternative reads ($r = N_B/(N_A + N_B)$), which theoretically falls into interval $[0, 1]$. When $r = 0$ or $r = 1$, there is expression of only one allele, which means this position is mono-allelically expressed; when $r \subset (0, 1)$, there

is expression from both two alleles on this position, meaning that this position is bi-allelically expressed; specifically when $r = 0.5$, the expression comes half from the reference allele and half from the alternative allele. As the key question is whether bi-allelic expression happens, the initial r value is converted to an absolute r value, with:

$$\text{absolute } r = \begin{cases} r, & \text{if } r \geqslant 0.5 \\ 1 - r, & \text{if } r < 0.5. \end{cases} \tag{2.4}$$

The absolute r falls into interval $[0.5, 1]$, while 0.5 stands for (balanced) bi-allelic expression and 1 stands for mono-allelic expression. The aIS of one h-iPSC line is defined as the average of absolute r on all heterozygous positions on X chromosome. Theoretically, the median of absolute r could also be used for the computation of aIS. The difference of these two measurement is that, when using median, the small number of positions where bi-allelic expression happens is ignored; in contrast, when using the average, all heterozygous positions are considered for their allelic expression level (figure 2.7).

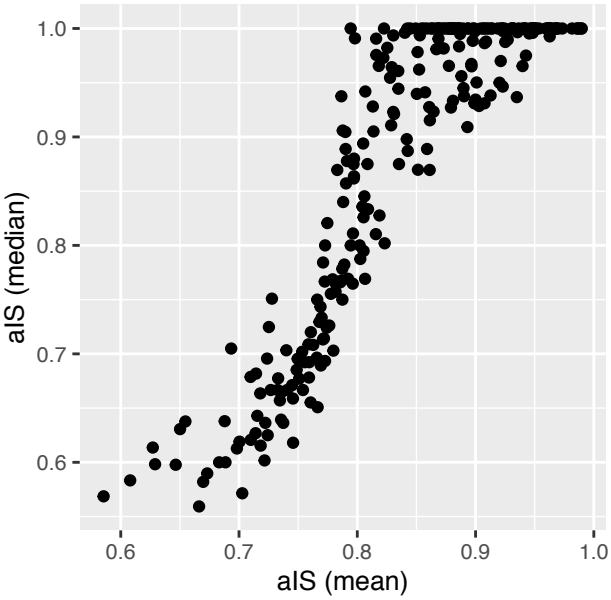


Figure 2.7: The comparison of aIS computed by the average or median. The median-based aIS ignores the small number of positions with bi-allelic expression while the average takes all positions into the consideration for the activity level of X chromosome.

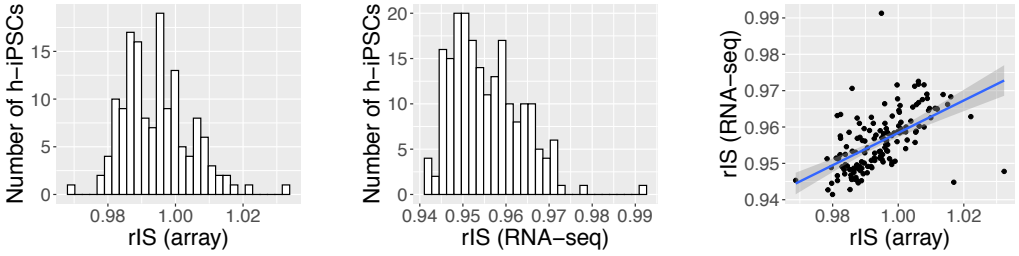
Since the motivation is to measure the activation level of the entire X chromosome, the average of absolute r is used as the aIS for female h-iPSCs for the association analysis between XCI metrics (section 2.4), as well as in further analysis of this thesis.

2.3.2 Definition and computation of rIS

The XCI is a sex-chromosome dosage compensation mechanism that equalizes the expression level of X chromosome in the two genders (Lyon 1961, Brockdorff et al. 2015, Heard et al. 1997, Avner et al. 2001). XCI loss is assumed to result in the increase of X-related expression level. With this assumption, the rIS is defined as the ratio of mean expression level of genes on the X chromosome over the mean expression level of genes on the autosomes, with formula:

$$\text{rIS} = \frac{\text{mean (expression of X chromosome)}}{\text{mean (expression of autosomes)}}. \tag{2.5}$$

The microarray data were available for 148 independent female h-iPSCs lines. Figure 2.8 shows the distribution of rIS computed by the RNA-sequencing data and the microarray data (a and b), as well as the association of rIS computed by these two technologies for 148 overlapping lines (c, Pearson correlation = 0.56, adjusted R^2 in univariate linear regression = 0.31).



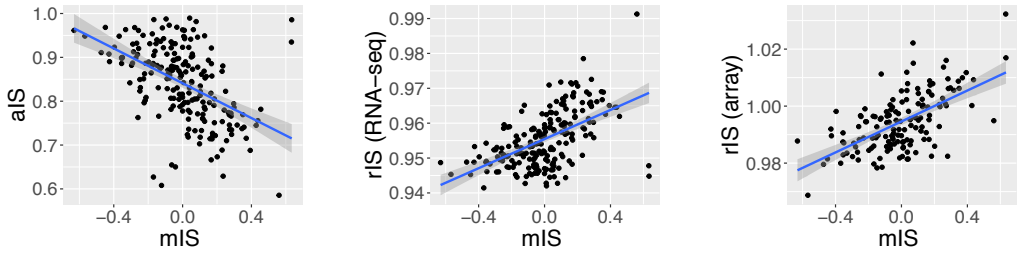
(a) (b) (c)

Figure 2.8: The rIS value for female h-iPSC lines in HipSci. a. Computed by microarray data. b. Computed by RNA-sequencing data. c. The association between rIS computed by two technologies for the overlap 148 lines (Pearson correlation = 0.56, adjusted R^2 in univariate linear regression = 0.31).

2.3.3 Biological conclusions

As mentioned in section 2.2, mIS is used as the standard representation of XCI. Here, I present the association between two expression based metrics and the mIS. For this analysis, I randomly selected one line per donor to remove donor effect, making a data set of 205 independent female h-iPSCs.
Figure 2.9 shows the association between the two expression-based XCI metrics and the mIS. By definition, aIS measures the ratio of bi-allelic expression on the X chromosome. A higher aIS value refers to more mono-allelic expression (proper XCI), while a smaller value refers to more bi-allelic expression (XCI loss). The association between mIS and aIS is consistent with their definition, with Pearson correlation value -0.50 (figure 2.9 a).

Positive correlation between mIS and rIS is observed in figure 2.9 (b and c) for



(a) (b) (c)

Figure 2.9: The association between mIS and expression-based XCI metric. a. Association between mIS and aIS for 205 female h-iPSCs. b. Association between mIS and RNA-seq-computed rIS for 205 female h-iPSCs. c. Association between mIS and array-computed rIS for 148 female h-iPSCs.

rIS using either RNA-sequencing data or microarray data (Pearson correlation = 0.55 and Pearson correlation = 0.52, respectively). This observed association is consistent with the assumption that, the loss of XCI will result in more expression on the X chromosome (specifically, X_i), proving the availability to use expression ratio as XCI metrics.

To conclude, the expression level of h-iPSCs is able to represent the XCI status in h-iPSC lines. For this usage, either the ratio of bi-allelic expressed genes on the X chromosome, or the ratio of average gene expression on the X chromosome over the average gene expression on the autosomes can be used. This result will help to maximize the utility of data in laboratories and research groups using female h-iPSC lines.

2.4 Similar XCI level in single cells of h-iPSC line joxm_1

Single-cell sequencing technology has greatly helped the cell-type classification, the cell-lineage analysis, as well as studies of disease-related genomics since its initial application (Gawad et al. 2016, Kalisky et al. 2011, Navin et al. 2011, De Bari et al. 2006, Gawad et al. 2014). In this section, I investigate the following question: What is XCI status at single cell level - do cells display a similar or a dispersed XCI status?

With the single-cell profiling by Linker et al. 2019, I present that for female line joxm_1, XCI patterns in 84 single cells are similar as in bulk level, using expression-based XCI metrics aIS.

2.4.1 Single cell data processing

The female h-iPSC line joxm_1 in HipSci (Kilpinen et al. 2017) shows an intermediate XCI loss level (aIS = 0.80; rank in 205 female h-iPSC lines: 70). Meanwhile, there is no sibling-line of joxm_1 in HipSci. Linker et al. 2019 generated a set of 84 cells from joxm_1, assayed by Smart-Seq2 (Picelli et al. 2014). Raw RNA-sequencing data in fastq format of these 84 samples are available on EMBL-EBI website with study ID PRJEB15062 (EMBL-EBI).

Data processing

The alignment of fastq files and variant calling were done by Dr. Angela Goncalves. The alignment of fastq files used Genome Reference Consortium Human Build 37 (GRCh37) as human genome reference, based on ENSEMBL version 74, downloaded from the ENSEMBL website (ENSEMBL 2010). The alignment used Kallisto version 0.43.0 (Bray et al. 2016). The variant calling process was carried out with the function *mpileup* of samtools, version 1.9 (Heng Li, B. Handsaker, et al. 2009), with minimum base quality more than 20. Heterozygous positions were extracted from exome sequencing of h-iPSC line joxm_1, which is the same as for computation of aIS using bulk RNA-sequencing data. For single cell samples, VCF files generated from variant call process contain 5000 variants on average, among which around 300 are non-zero variants.

2.4.2 Distribution of XCI at single cell level

With the definition of aIS in section 2.3.1, the average of absolute r value on X chromosome was used as aIS for single cell samples. The distribution of aIS in 84 single cell samples is shown in figure 2.10.

Figure 2.10 reveals that most cells of joxm_1 maintained similar XCI pattern as found at bulk level (50% cells with aIS $>$ 0.70). With the 84 cells profiled by Linker et al. 2019, a similar pattern of XCI is observed for the female h-iPSC line joxm_1. This is an inspiring result for the research of XCI because most studies investigated the XCI level with a large number of cells. Meanwhile, since this observation is based on one single line, I am very looking forward to further studies which explore the XCI pattern and the XCI process at single cell level with a larger sample size.

2.5 Predictor genes do not serve better than expression matrix for XCI representation

In section 2.3.3, I present that the expression level of h-iPSCs can be used as XCI metrics. The next question is, instead of using the entire expression level, is possible to use a couple of genes to estimate the XCI level in h-iPSCs?

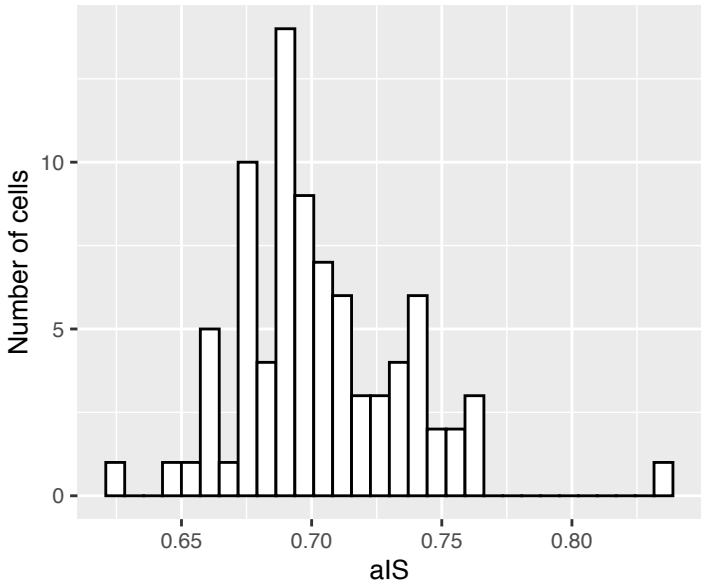


Figure 2.10: Histogram of aIS for 84 single cell samples of female h-iPSC joxm_1 (mean = 0.70).

When talking about marker gene(s) of XCI level, XIST, the key regulator of XCI, was widely used in previous studies as a single marker of XCI level and there is a lack of study to explore the predictive performance of XCI by a group of genes (C. J. Brown, Hendrich, et al. 1992, R. Brown et al. 1993, Anguera et al. 2012, Mekhoubad et al. 2012).

The major motivation of this section is to get a balance between using the entire expression level and using single-gene expression level for the representation of XCI: to maintain a similar predictive performance as the expression level and to avoid the bias of using a single predictor. To limit the number of genes used for the estimation of XCI level can also reduce the work load for other laboratories using h-iPSCs for scientific researches.

In this section, I used lasso regression, together with nested cross validation method for the selection of marker genes (Tibshirani 1996). In the predictive model, genome-wide expression level and X-linked expression level were used respectively as data input for marker gene selection. The technical details of the model and the predictive results are presented in section 2.5.1. The PEER correction method, which is a joint Bayesian framework, is applied to remove both known and unknown batch factors in the expression level (Stegle, Parts, Durbin, et al. 2010, Stegle, Parts, Piipari, et al. 2012).

The computational analysis and results visualization in this section was conducted in R, version 3.4.0 (R Core Team 2017), with package glmnet (Friedman

et al. 2010), package caret (Jed Wing et al. 2018), package ggplot2 (Wickham 2016) and package PEER (version 1.3, Stegle, Parts, Piipari, et al. 2012).

2.5.1 Using lasso regression and nested cross-validation for selection of predictor genes

For predictive modelling with high dimensional data input, the penalized linear regression is a widely used method, since it controls the number of variables in the final model by minimizing the penalized residual sum of squares (Hastie et al. 2009). The lasso regression is one of the most widely used penalized linear model because it sets the coefficients to exactly zero when variables are not relevant in the predictive model, thus only a small number of truly important variables among the enormous data input are included in the model (Hastie et al. 2009, J. Fan et al. 2010, Kyung et al. 2010).

The Lagrangian form of lasso regression (Hastie et al. 2009) is written as formula 2.6 :

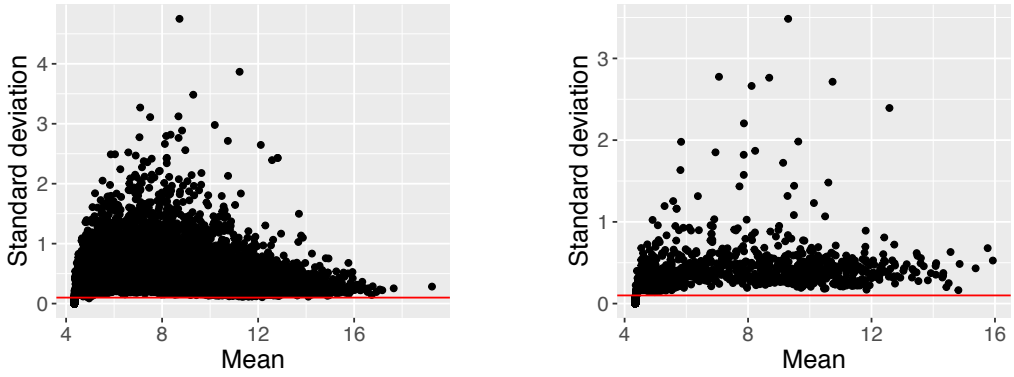
$$\hat{\beta}^{lasso} = argmin_{\beta}(\frac{1}{2}\sum_{i=1}^{N}(y_1 - \beta_0 - \sum_{j=1}^{P} x_{ij}\beta_j)^2 + \lambda\sum_{j=1}^{p}|\beta_j|). \tag{2.6}$$

The lasso regression controls the size of predictors by the tuning parameter λ, which was set to $1,000$ in the analysis of this chapter. Since the lasso regression is able to select a small number of variables from a large data input, it is important to know the predictive performance of these variables in a different, unseen data set. For this aim, the nested cross validation was applied in combination with the lasso regression.

In brief, the nested cross validation is based on data split of the original data set and is a recurrent model-fitting and model-validating process. To describe it in details, the entire data set was split into 10 folds in a random manner with a fixed initial seed in R (R Core Team 2017), labeled as fold 1, fold 2, ... fold 10. While the fold $k(k \in [1, 10])$ was used as the test set, the rest 9 folds were used as the train set. In the train set, an internal 10-fold cross validation was applied, so that a best model was selected. This best model was then validated on the test set fold k and the predictive performance was measured with the RSS value. The test fold moved from fold 1 to fold 10, with each fold being used only once as the test set. The entire validation result makes it possible to estimate the predictive performance of the lasso regression in the whole data set, meanwhile each validation result was obtained when this sample was not seen by the training model.

To control the donor effect, a random selection of one line per donor was executed, making a data set of in total 205 female h-iPSCs. The expression level of these 205 h-iPSCs was used for gene-marker selection, including 2,190 X-located genes and 52,220 autosomal genes. Before fitting the predictive model, a filtering process was carried out to remove genes which standard deviation is smaller

than 0.1 (figure 2.11).



(a) (b)

Figure 2.11: The standard deviation and the average of gene expression in 205 female h-iPSCs (random selection one line per donor). a. Genes on all chromosomes. b. Genes on the X chromosome (red horizontal line: standard deviation = 0.1)

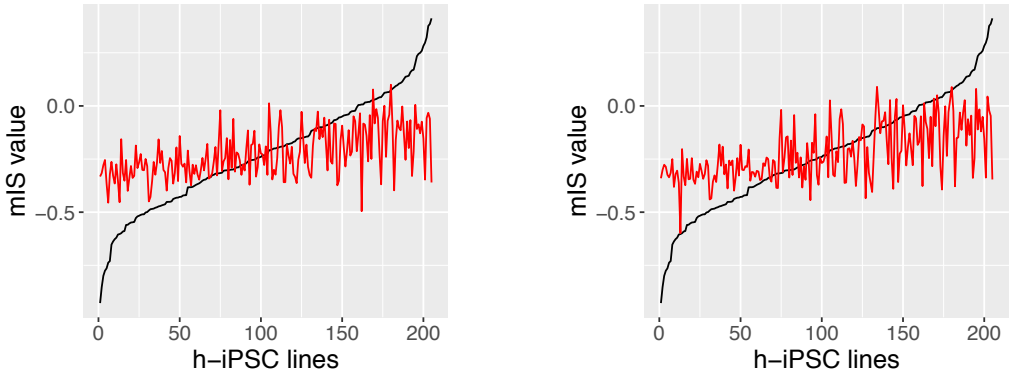
Figure 2.11 shows that genes with standard deviation smaller than 0.1 also tend to have low average value in the entire data set (average expression value < 5), meanwhile, most of these genes are X-located. Therefore, removing genes with standard deviation smaller than 0.1 would remove X-located genes with small expression level and slight variation among female h-iPSCs, which are assumed to not have an important rule in the XCI status.

After the filtering, 39,856 autosomal genes and 1,580 X-located genes remained, restricting the total number of genes to 41,436. The mIS value of these 205 female h-iPSCs was used to represent XCI level and was included in the lasso regression as the dependent variable.

Figure 2.12 presents the prediction of lasso regression for 205 h-iPSCs. At the same time, it helps us to visualize the different predictive performance on different h-iPSC samples: the prediction for h-iPSCs with either very low or very high mIS values was not as good as the prediction for those with intermediate mIS values. With the lasso regression, using either all genes or only X-located genes have similar predictive results: the RSS of the prediction by all genes was 9.91 while by X-located genes was 9.47.

2.5.2 Using PEER method for batch effect correction in expression level

Since the RNA-sequencing data in the HipSci project was generated by the single institute and went through the same processing procedures (Kilpinen et al. 2017 and section 1.1.1 of this thesis), here, major concern is to remove the known



(a) (b)

Figure 2.12: The prediction of mIS value for 205 female h-iPSCs using lasso method and nested cross validation (black: true mIS value; red: estimated mIS value). X-axis is the h-iPSC lines in the ascending order by their true mIS value and y-axis is numeric values. Each point stands for a true/estimated mIS for one h-iPSC line, meanwhile the estimated mIS for this line is obtained when it is not included in the train set in the nested cross validation. a. Using expression level from all chromosomes as data input (RSS = 9.91). b. Using expression level of X chromosome as data input (RSS = 9.47).

and unknown in-study batch factors in the expression data. As introduced before, the PEER method, which is abbreviation of probabilistic estimation of expression residuals, uses additive Bayesian network to infer hidden factors and their effects in gene expression matrix (Stegle, Parts, Durbin, et al. 2010, Stegle, Parts, Piipari, et al. 2012). In practice, PEER method uses general additive linear model, which contains three independent data (groups): the true expression, the known factors and the unknown factors, assuming that the gene expression is influenced in additive manner from different sources of batch factors (Stegle, Parts, Durbin, et al. 2010, Stegle, Parts, Piipari, et al. 2012). The Bayesian learning (Jordan et al. 1999) is used for the parameter in PEER, thus the estimation of one data group (i.e. known factors) would take all other parts of the model into account (Stegle, Parts, Durbin, et al. 2010). The output of PEER correction contains three parts: residuals, which is used as corrected expression level, weights of the inferred confounders and precision (the inverse variance) of the weights (Stegle, Parts, Piipari, et al. 2012).

The correction was executed on all 54,410 genes with RNA-sequencing data. As suggested by the tutorial of the PEER package (Stegle, Parts, Piipari, et al. 2012), the number of iterations was set to $1,000$ and the number of factors was initially set to 50, as the tutorial recommended to use 25% of the number of individuals as the initial number of factors (Stegle, Parts, Piipari, et al. 2012).

Figure 2.13 contains the standard plot-output of PEER package. It presents the the posterior variance of the factor weights and the situation of convergence:

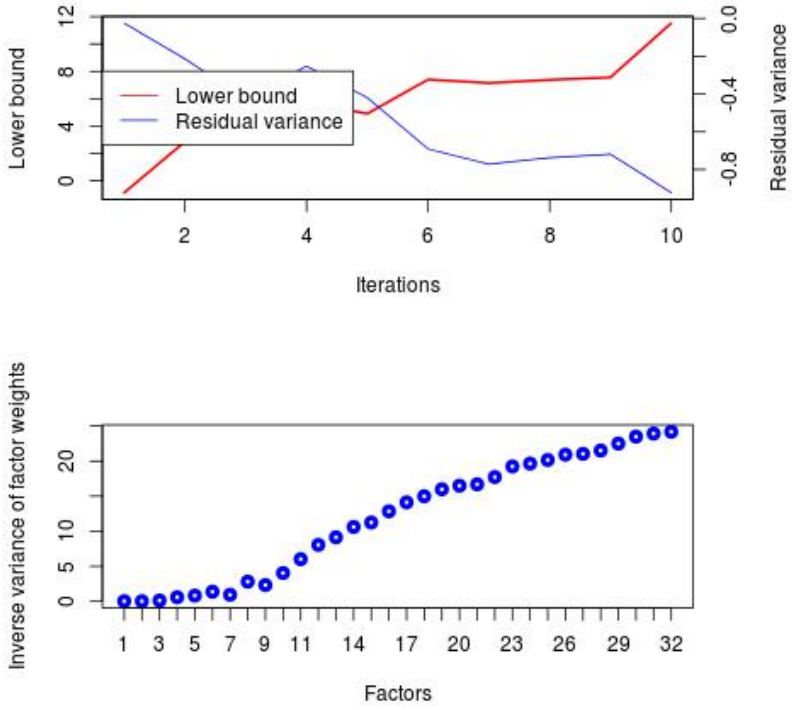
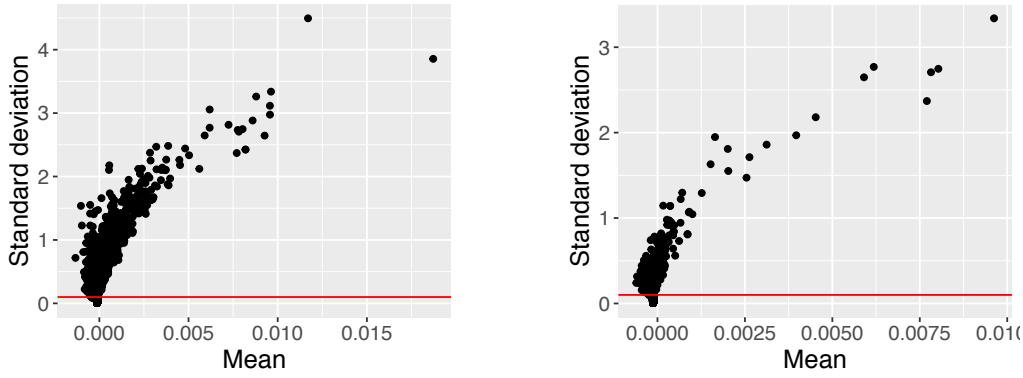


Figure 2.13: PEER correction results for RNA-sequencing data of all 54,410 genes. The inverse variance showed an 'elbow-like' change at the 9th factor, revealing the model achieved convergence at this factor.

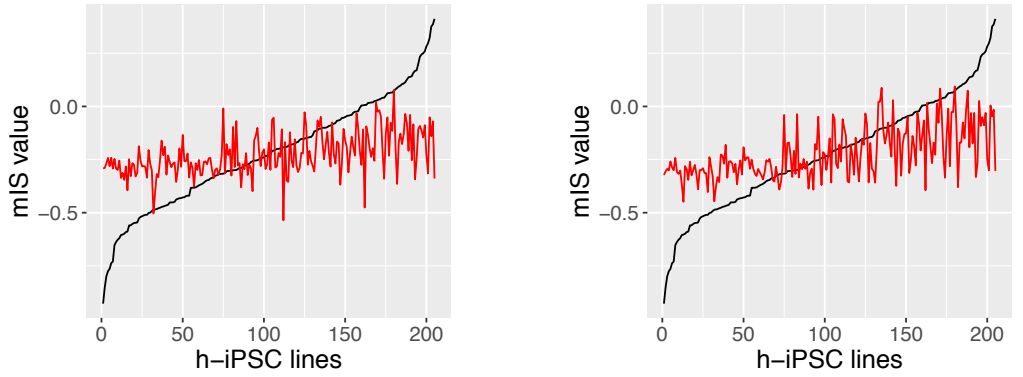
the inverse variance showed an 'elbow-like' change at the 9th factor, representing that the inverse variance had fluctuation before this factor while had constant increase afterwards, thus the model achieved convergence at this factor (Stegle, Parts, Piipari, et al. 2012). The average and the standard deviation of the corrected expression level is presented in figure 2.14, which shows a more concentrated mean-sd association compared to figure 2.11, meaning that the difference between expression levels is reduced after the correction.

The predictive performance of corrected gene expression level is shown in figure 2.15. The RSS of the prediction by PEER-corrected all-chromosome genes is 10.9 and is 9.52 by PEER-corrected X-located genes. By comparing RSS values, the new predictive results with PEER-corrected expression level are not significantly better than previous results (RSS = 9.91 and RSS =9.41 using uncorrected expression level of all genes and X-linked genes, respectively). The potential reason for this observation is that the PEER method is a very powerful correction method that it removes all factors which might result in the alteration of expression, including the XCI level. Therefore, it is unable to achieve a signifi-



(a) (b)

Figure 2.14: The standard deviation and the average of PEER-corrected gene expression in 205 female h-iPSCs. a. Genes on all chromosomes. b. Genes on the X chromosome (red horizontal line: standard deviation = 0.1)



(a) (b)

Figure 2.15: The prediction of mIS value for 205 female h-iPSCs using lasso regression and nested cross validation with PEER-corrected expression level (black: true mIS value; red: estimated mIS value). X-axis is the h-iPSC lines in the ascending order by their true mIS value and y-axis is numeric values. Each point stands for a true/estimated mIS for one h-iPSC line, meanwhile the estimated mIS for this line is obtained when it is not included in the train set in the nested cross validation. a. Using corrected expression level from all chromosomes as data input (RSS = 10.9). b. Using corrected expression level of X chromosome as data input (RSS = 9.52)

cant improvement of predictions.

To conclude, with and without batch effect corrected, the group of genes selected by lasso regression show a mild predictive performance for mIS, especially for h-iPSC lines with very low or very high mIS values. However, since that the very low or very high mIS referring to a respectively proper XCI or complete

XCI loss in female h-iPSCs, a non-accurate prediction of XCI level for these lines is not helpful for their usage, particularly in disease modeling. Therefore, it is recommended to use the two expression-based metrics, namely aIS and rIS, for the XCI estimation (section 2.3).

2.6 Discussion: the XCI heterogeneity in female h-iPSCs

General conclusion

This chapter presents XCI heterogeneity in female h-iPSCs at population level with DNA methylation level of the X chromosome and shows the utility of expression level for the representation of XCI status in h-iPSC lines.

Using 273 female h-iPSCs from HipSci (Kilpinen et al. 2017), the analysis was executed with the data set from the single institute, with similar experimental protocol and screening process for all h-iPSCs, which guaranteed a low-bias initial data set. Using the methylation level, I demonstrate that XCI heterogeneity exists in this large-scale h-iPSC data set, showing as 1% h-iPSCs (4 out of 273 h-iPSCs) display a complete XCI loss while the majority has a variable XCI level. Meanwhile, h-iPSCs generated from the same donor have more similar XCI level to one another than h-iPSCs from different donors.

The importance to include XCI variation in studies using female h-iPSCs

The XCI level is a genetic marker in h-iPSCs and has been found associated with cell development, cellular functions and human diseases (R. Brown et al. 1993, Hao Wu et al. 2014, S. Wang et al. 2013, Brix et al. 2005, Santiwatana et al. 2018). Thus, when use h-iPSCs for disease modeling, cell therapy or drug development, XCI variation might lead to alteration in genomics, transcriptomics, or other downstream phenotypes. Another important point is that, in human, the X chromosome contains the largest groups of immune related genes (Libert et al. 2010), hence the XCI heterogeneity may lead to different immune activity of h-iPSC lines. For these reasons, it is important to include the XCI level of h-iPSCs as a covariate in biological and clinical researches.

Facilitation for other researchers: expression based XCI metrics

It is common to have expression level other than methylation level of h-iPSCs in research groups who study h-iPSCs, since the expression level is required for the pluripotency test, which is widely used during the generation of h-iPSC lines (Müller et al. 2011). Considering this situation, this thesis introduced two expression-based XCI metrics and demonstrated that they correlate well with the methylation-based XCI metrics.

The inspiration for the following analysis

Many questions rose up following the demonstration of XCI heterogeneity. What are sources of this variation? Is it possible to control the variation by experimental settings? Do older females have more abnormal XCI status when compared to younger females? What are the consequences of this heterogeneity? In this thesis, these questions are studied, discussed and compared with previous observations. Based on HipSci (Kilpinen et al. 2017), the contribution of this thesis to this research field is the presentation of the XCI heterogeneity in female h-iPSCs and to answer these key questions concerning XCI heterogeneity.

Chapter 3

Sources of XCI heterogeneity in female h-iPSCs

- What are the most important sources of XCI heterogeneity?
The donor effect and XIST, while XIST is not a perfect marker.

The observation of line-to-line variability of XCI in female h-iPSCs raises an important question: what are sources of this heterogeneity?

This question is also a key concern for scientists working with h-iPSCs since the control of XCI heterogeneity will largely help the usage of h-iPSCs in disease modeling and in drug development (Galupa et al. 2018, Schöndorf et al. 2014, Y.-T. Lin et al. 2018). Below, I summarize potential sources in the 'life-stage' of h-iPSCs, as well as previous studies regarding their effects on XCI level in h-iPSCs.

Donor metadata

Donor is the root of h-iPSCs, therefore the donor information is an interesting factor for the study of variation in h-iPSCs. Previous studies found a negative effect of donor age in the induction of h-iPSCs from fibroblasts and age-related DNA methylation at some CpG sites in h-iPSCs from blood cells (Trokovic et al. 2015, Mackey et al. 2018, Sardo et al. 2017, Mahmoudi et al. 2012, Mertens et al. 2018). In HipSci (Kilpinen et al. 2017), the age and the health status of donors were collected, which allow the investigation of whether the XCI variation is associated with the increase of age and whether the XCI variation is different in h-iPSCs from healthy donors or from donors with genetic diseases (Bardet-Biedl syndrome and neonatal diabetes).

Cell culture factors

The h-iPSC lines are cultured in cell culture media for a certain of passages before usage. An observation of XCI loss and DNA methylation in h-iPSCs with prolonged cell culture time was reported by several studies (Mekhoubad

et al. 2012, Trokovic et al. 2015, Anguera et al. 2012, Nazor et al. 2012). Previous studies had limits in following aspects: the number of h-iPSC lines was limited ($n = 11$ in Trokovic et al. 2015 and $n = 12$ in Mekhoubad et al. 2012); XIST expression level was used as representation of XCI status (Mekhoubad et al. 2012, Anguera et al. 2012) and culture time was limited to very early passages (passage 0-7, Anguera et al. 2012). HipSci (Kilpinen et al. 2017) applied much longer and various culture time for h-iPSCs (min = 24 days, max = 240 days, details in section 1.1), therefore this thesis is able to study the long-time-effect in XCI variation in a large-scale data set of h-iPSCs. Besides, two media (Feeder Free and Feeder Dependent, below FF and FD) were used for cell culture. Kilpinen et al. 2017 reported the stratification of pluripotency score due to cell culture media in 711 h-iPSCs in HipSci. Therefore, this thesis investigates specifically the media effect in XCI-heterogeneity for female h-iPSCs.

The key biological regulator, XIST

The long non-coding RNA XIST is the key regulation factor for the start of the XCI process and has been intensively studied for its role in h-iPSCs (Lyon 1961, Penny et al. 1996, C. J. Brown, Hendrich, et al. 1992, Avner et al. 2001, Galupa et al. 2018, Mekhoubad et al. 2012, Anguera et al. 2012). This thesis uses population-size data set (Kilpinen et al. 2017) for the association study between XIST expression and the XCI status in female h-iPSCs, as well as the time-effect in XIST expression.

Genetic alterations during generation

Copy number alterations between h-iPSCs and their progenitor cells were observed by previous studies (Laurent et al. 2011, Amps et al. 2011, Abyzov et al. 2012). In HipSci, Kilpinen et al. 2017 also reported genetic alterations between h-iPSCs and progenitor cells (fibroblasts), including both trisomy of X chromosome and copy number alterations (CNAs) on all chromosomes. This chapter summarizes the CNA level in female h-iPSCs and investigates the association between recurrent CNAs with XCI level in female h-iPSCs. Here, I firstly present the association between these potential sources and XCI level in female h-iPSCs separately, then apply variance component analyses (VCAs) to investigate their contributions to the XCI variability.

3.1 Strong effect of cell culture media and light effect of culture time in XCI variation

Experimental factors can bring bias into induction rate, pluripotency level and functions of h-iPSCs, thus to minimize batch variables is a key concern in the field (Mekhoubad et al. 2012, Kilpinen et al. 2017, G. Chen et al. 2011, Liang et al. 2013).

For 273 female h-iPSCs in the analysis, 222 lines were cultured in media FF and 51 lines cultured in media FD. A strong stratification of XCI level by culture media was observed (p-value $< 1.8 \times 10^{-12}$ for mIS and p-value $< 3.0 \times 10^{-12}$ for aIS, figure 3.1 a and b). With two XCI metrics, it is observed that h-iPSC lines which were cultured in media FD display stronger XCI loss, showing as higher mIS and lower aIS, when compared to lines cultured in media FF. Therefore, in further analysis, the media effect is accounted either by including it as covariate or by stratification.

Stratified by media, the XCI level is not associated with cell culture time in the long cell culture up to 240 days, shown with two XCI metrics mIS and aIS ($p > 0.3$, figure 3.1 c and d). Meanwhile, for h-iPSCs with relatively short culture time (time < 50 days, 94 lines, all cultured with media FF), a slight erosion of XCI is observed with prolongation of time, showing as the decrease of aIS with culture time ($p = 0.09$, figure 3.2).

This result reveals that XCI status is not associated with cell culture time in long term. Meanwhile, within relative short culture time ($<$ 50 days in HipSci), a slight loss of XCI is accompanied with longer culture. 50 days of cell culture refers to approximate passage 15 for lines cultured in media FF (all 94 h-iPSC lines). Mekhoubad et al. 2012 observed an erosion of XCI in h-iPSCs from low-passge (passage 5-6) to high-passage (passage 19-24) by loss of XCI markers (XIST cloud and H3-K27 tri-methylation). Briggs et al. 2015 also observed the loss of XIST at late-passage (passage 20) compared to the initial stage (passage 0) at single cell level. To conclude the slight loss of XCI is expected with prolongation of culture time in early-middle passages, however the XCI level maintains the similar level when the culture time is significantly extended.

3.2 Donor age and health condition do not show association with XCI variation

In HipSci (Kilpinen et al. 2017), the age information was recorded for 200 out of 205 female donors. Among these female donors, the youngest was 5-year-old while the oldest was 80-year-old. The average and median of age for female donors were both 55. The collection for health information for female donors was successfully done on all 205 female donors. Even though most donors enrolled in this project were under general healthy situation (referred as 'normal', number = 170), 21 donors had Bardet-Biedl syndrome (bbs) and 14 donors had neonatal diabetes (nd).

To test whether the age and health condition of female donor effect XCI level, a linear regression is fit between mIS and donor's age or donor's health condition, respectively, where no association is found between mIS and these two donor meta factors (figure 3.3).

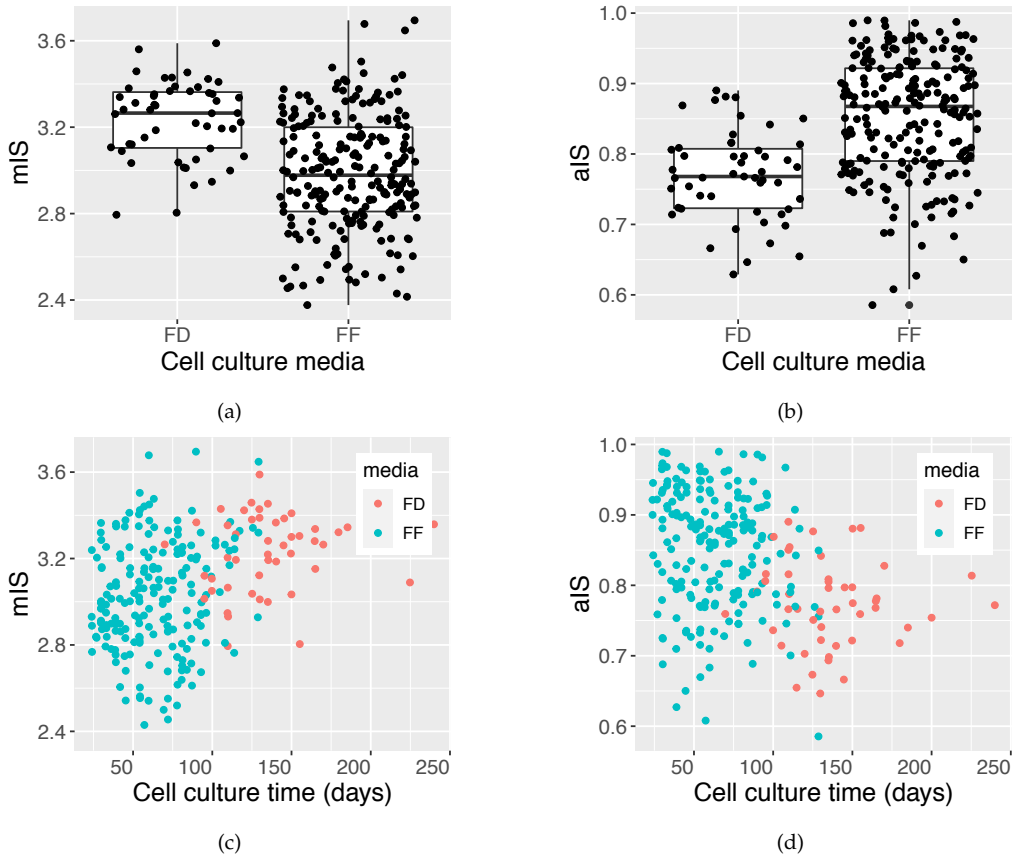


Figure 3.1: Effects of cell culture on XCI-heterogeneity in female h-iPSCs using XCI metrics mIS and aIS. a and b. Cell culture media has strong stratification effect on XCI level, showing with mIS (a, p-value $< 1.8 \times 10^{-12}$) and aIS (b, p-value $< 3.0 \times 10^{-12}$). c and d. Stratified by culture media, there is no association between cell culture time and mIS (c) or aIS (d, p-value > 0.3).

This result reveals that the XCI heterogeneity widely exists in h-iPSCs, regardless of age and health condition of donors. For further researches with female h-iPSCs, I recommend scientists to include XCI level in the study of h-iPSCs regardless of the recruitment of donors.

3.3 The association and gap between XCI loss and XIST expression

The long coding RNA XIST is a key factor in the establishment of XCI in mammals (Penny et al. 1996, C. J. Brown, Hendrich, et al. 1992, Berg et al. 2009, Avner et al. 2001). In human, XIST locates on both two X chromosomes and is expressed specifically on the inactive X chromosome (Xi), which leads to the silencing of Xi (C. J. Brown, Hendrich, et al. 1992, Avner et al. 2001, Wutz et al. 2000, Galupa et al. 2018). Because of its XCI-related function, XIST has been

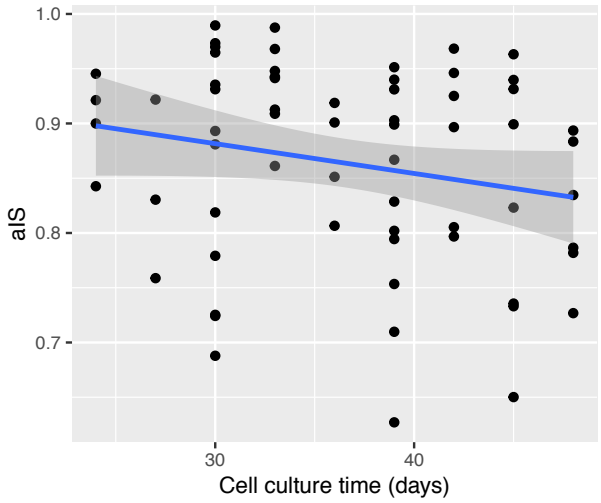
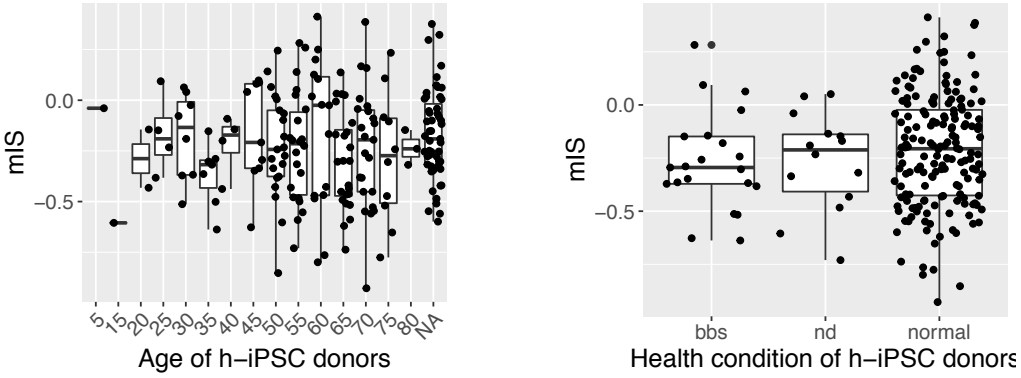


Figure 3.2: For 94 h-iPSC lines with relatively short culture time (< 50 days), a slight loss of XCI is observed with the increase of culture time (p-value = 0.09)



(a) (b)

Figure 3.3: The XCI variation exists in h-iPSCs generated from different age (a) and different health conditions (b) of donors. The mIS is corrected by donor effect, cell culture media and technical factors in the methylation array. The age of donors is shown in 5-year interval (a). The abbreviations of health conditions: 'bbs' stands for Bardet-Biedl syndrome, 'nd' stands for neonatal diabetes and 'normal' stands for healthy donors.

used as an XCI marker and was intensively studied for its association with structural, genetics and disease-related characters of h-iPSCs (Mekhoubad et al. 2012, Briggs et al. 2015, Ananiev et al. 2011, Splinter et al. 2011).

This section presents the general picture of XIST expression in h-iPSCs from HipSci (Kilpinen et al. 2017): the prevalence of XIST expression in h-iPSCs of both genders; the cell culture effects on XIST expression and the association between XIST expression and XCI level.

3.3.1 On and off mode of XIST in female h-iPSCs

XIST expression level is available for all 273 female and all 219 male h-iPSCs. In female h-iPSCs, a clear bimodal distribution of XIST expression is observed, compared to the uniform expression level in male h-iPSCs.

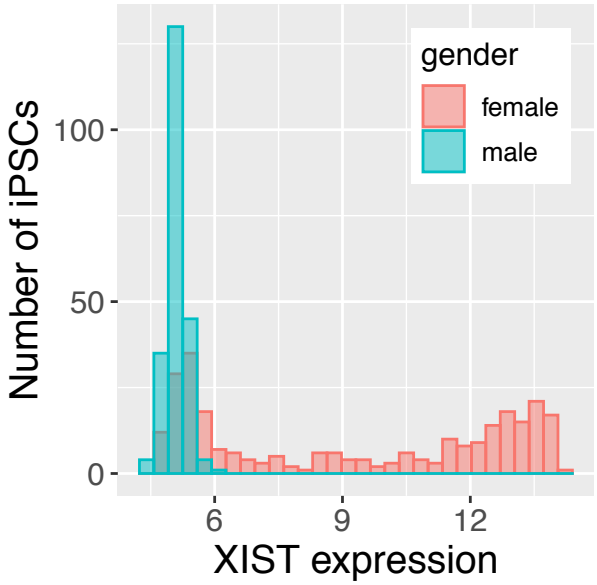


Figure 3.4: XIST expression displays bimodal distribution in 273 female h-iPSCs, compared with the unimodal expression level in 219 male h-iPSCs.

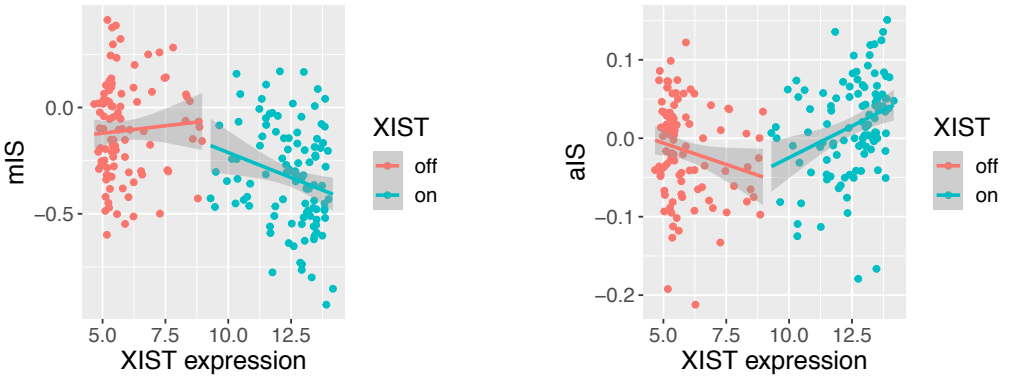
XIST expression level exhibits a clear bimodal distribution in 273 female h-iPSCs, named as on/off mode (figure 3.4). The cluster of female h-iPSCs with 'on-mode' have an average XIST expression at 12.5 and the cluster of lines with 'off-mode' have an average expression level at 5.9, which is similar to the XIST expression level in 219 male h-iPSCs (mean = 5.9, standard deviation = 0.24, figure 3.4).

Since that XCI only takes place in cells with multiple X chromosomes (C. J. Brown, Ballabio, et al. 1991), XIST is not expressed in male h-iPSCs, thus the 'off-mode' in female h-iPSCs actually referring to the non-expression of XIST.

This result reveals that at population level, female h-iPSCs can display two modes of XIST expression. In section 3.3.2 and 3.3.3, I present the association between XCI level and XIST expression under these two modes and the experimental factor which is directly associated with the mode of XIST.

3.3.2 Association between XIST and XCI variation

Using k-means clustering method (Lloyd 1982, MacQueen et al. 1967), with k = 2, 106 female h-iPSCs are grouped as on-mode, while 99 female h-iPSCs are grouped as off-mode. XCI level is clearly stratified by the XIST mode: using either mIS or aIS as XCI metric, h-iPSC lines with on-mode XIST display a more proper XCI level, showing as lower mIS and higher aIS (p-value < 0.01 for both XCI metrics). Figure 3.5 shows the association of XIST expression with two XCI metrics, mIS (a) and aIS (b), with female h-iPSCs colored by the mode of XIST.



(a) (b)

Figure 3.5: The distribution of XCI level on on/off of XIST. a. Using mIS as XCI metrics, which is corrected with donor effect, cell culture media and technical factors in the methylation array. b. Using aIS as XCI metrics, which is corrected by donor effect and cell culture media.

The association analysis between XIST expression and XCI variation was stratified by the 'on/off' mode of XIST. Female h-iPSCs with off-mode XIST display a large variation of XCI level, from almost proper XCI to complete XCI loss, shown with both mIS and aIS (red points in figure 3.5). On the other hand, for female h-iPSCs with on-mode XIST at on-mode (mean = 12.5), there is a clear association between XIST expression and XCI level: the Pearson correlation between XIST expression and mIS equals to -0.24 (p-value = 0.01, figure 3.5 a) and the Pearson correlation between XIST expression and aIS equals to 0.31 (p-value = 0.01, figure 3.5 b). When only looking into lines which are cultured with media FF, the association between on-mode XIST and aIS is stronger: the Pearson correlation equals to 0.51 (p-value $= 5.8 \times 10^{-8}$, figure 3.6).

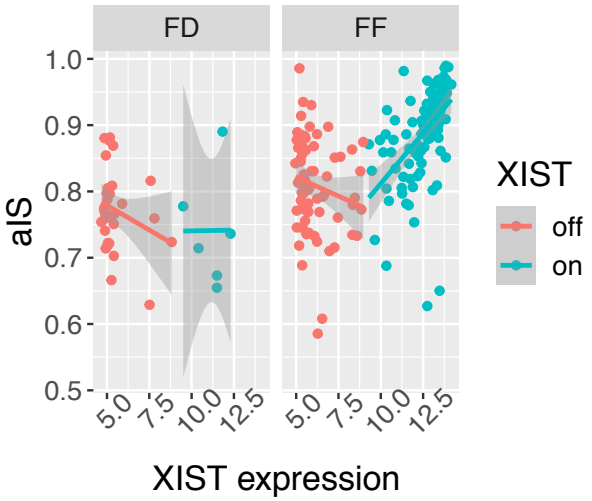


Figure 3.6: The distribution of raw aIS value on XIST, stratified by the cell culture media.

With current results, a number of h-iPSCs with on-mode XIST still display XCI loss (cyan points with intermediate or low mIS/aIS value in figure 3.5). Meanwhile, h-iPSCs with off-mode XIST show a wide range of XCI level, shown as various methylation and bi-allelic expression level (red points in 3.5). This result indicates that XIST is not a perfect marker of XCI, regardless of its important role in the initialization of the XCI process.

Some lines can be observed with intermediate level of XIST expression in figure 3.5. To have a detailed look at these lines, another k-means clustering was applied with k = 3, resulting in 82 female h-iPSCs grouped as on-mode (mean = 13.0), 90 grouped as off-mode (mean = 5.6) while 33 grouped as middle (mean = 10.1), shown in figure 3.7. Using mIS as XCI metrics, the association between XCI and XIST expression is most clear in h-iPSCs in on-mode group (Pearson correlation = -0.14), slightly less clear in h-iPSCs in off-mode group (Pearson correlation = 0.09) and not observed in h-iPSCs in the middle (Pearson correlation = -0.06).

Since the result based on the new grouping is consistent with the previous conclusion in general, to keep information brief, in the following analysis, the expression level of XIST is always described as on/off mode, using clustering result with k = 2.

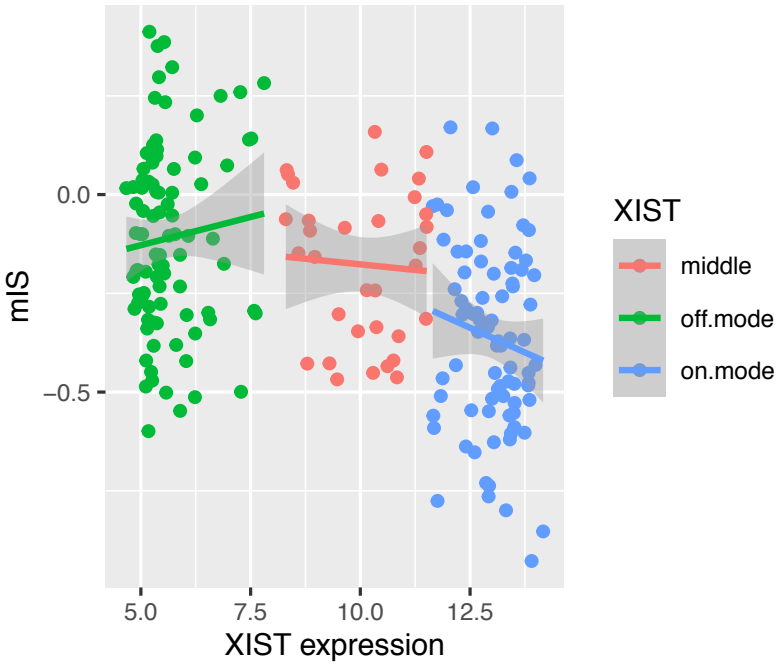


Figure 3.7: The distribution of XCI level with XIST in three clusters (on-mode: 82 h-iPSCs, average XIST expression = 13.0; middle: 33 h-iPSCs, average XIST expression = 10.1; off-mode: 90 h-iPSCs, average XIST expression = 5.6)

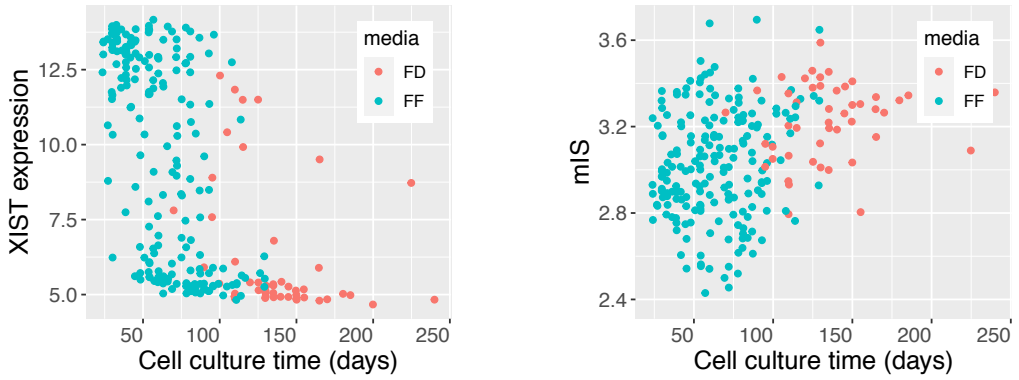
3.3.3 XIST expression sharply drops with culture time while not accompanied by alteration in XCI level

Considering the important role of XIST in XCI, I investigated whether the donor meta data or cell culture settings had an effect in XIST expression level in female h-iPSCs. The most interesting observation is the time effect in XIST expression: at approximately day 50 in cell culture, the XIST expression drops sharply from 'on-mode' (mean = 12.5) to 'off-mode' (average = 5.5), while is not accompanied by an alteration in XCI level in female h-iPSCs (figure 3.8).

In HipSci (Kilpinen et al. 2017), all lines cultured shorter than 50 days were cultured in medium FF, which means that cell culture media is not a confounding factor of the drop of XIST expression level.

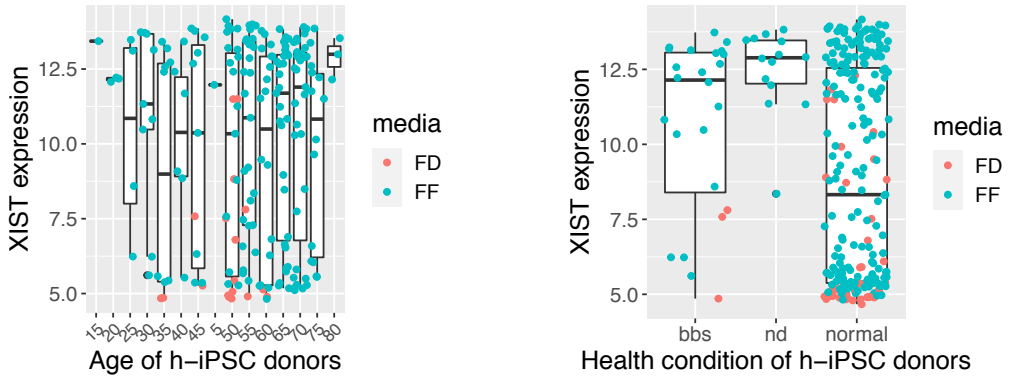
With the exception of cell culture time, XIST expression does not show association with donor age or donor health condition: female h-iPSCs of both 'on-mode' and 'off-mode' can be found in any group of these two factors with more than four lines (figure 3.9).

This section shows the bimodal distribution of XIST expression in 273 female h-iPSCs and presents the association between XIST expression and XCI level. Several studies reported similar observation of XIST expression loss with culture:



(a) (b)

Figure 3.8: The effect of cell culture time for female h-iPSCs. a. XIST expression drops sharply at approximately day 50. b. mIS does not change significantly with cell culture time, stratified by cell culture media.



(a) (b)

Figure 3.9: Similar XIST expression in age and health conditions of female donors. a. For age groups with with more than four h-iPSCs, XIST of both modes can be observed. b. Donors with neonatal diabetes (nd) show high and intermediate XIST expression level while donors with Bardet-Biedl syndrom (bbs) and healthy donors (normal) show clear high and low XIST expression level.

Geens et al. 2016 using 10 human pluripotent stem cells (both h-iPSC and h-ES cells); Mekhoubad et al. 2012 using h-iPSCs generated from fibroblasts from donors with Lesch-Nyhan syndrome and Briggs et al. 2015 presenting XIST expression at single cell level. However, the observation that female h-iPSCs with low XIST expression level exhibits various XCI level was unexpected. Considering these results, I assume that the culture time might change the interaction pattern between XIST expression and the XCI process. Furthermore, the gap between XIST expression and the XCI process, shown as reactivation of X-linked genes, were observed and reported by Vallot (Vallot, Huret, et al. 2013, Vallot, Ouimette, et al. 2015, Vallot, Patrat, et al. 2017), indicating that XIST expression is a relatively poor marker of XCI level, regardless of its critical role to trigger the XCI process.

3.4 Recurrent genetic alterations are not associated with XCI variation

HipSci (Kilpinen et al. 2017) used genotyping array to detect copy number alterations (CNAs) between h-iPSCs and their progenitor fibroblasts, defined as genetic abnormalities of over 200 kilobase (kb) occurring in at least 20% of the cells (technical details in section 1.1.1). For 711 h-iPSCs of both genders reported by Kilpinen et al. 2017, 41% of generated lines and 18% of selected lines contain one or more CNAs, including whole-chromosome duplication of the X chromosome (hereafter, trisomy of X chromosome), duplications and deletions on sub-chromosomal regions (hereafter, CNAs). Even though genetic alterations in h-iPSCs were also reported by previous studies, there is a lack of study for the association between these alterations and XCI level in h-iPSCs.

3.4.1 Trisomy of X chromosome

Among 273 female h-iPSCs, 14 lines (5.1%) exhibit trisomy of X chromosome, whereas the maximum number of X chromosome copy is 3.12. The general distribution of XCI metrics on X chromosome trisomy is shown in figure 3.10 (a and b). Even though there is a statistical association between X chromosome trisomy and XCI level in general (p-value < 0.01 for both XCI metrics), this association is confounded by the XIST mode: among 14 lines with X chromosome trisomy, 12 of them have 'on-mode' XIST (figure 3.10 c). When looking into h-iPSC lines with 'on-mode' XIST, XCI level is not associated with X chromosome trisomy ($p > 0.2$ for both XCI metrics). Furthermore, the vast majority of lines with X chromosome trisomy (12 out of 14) were generated from healthy donors, thus X chromosome trisomy is not a disease-related alteration in the induction of h-iPSCs (figure 3.10 d).

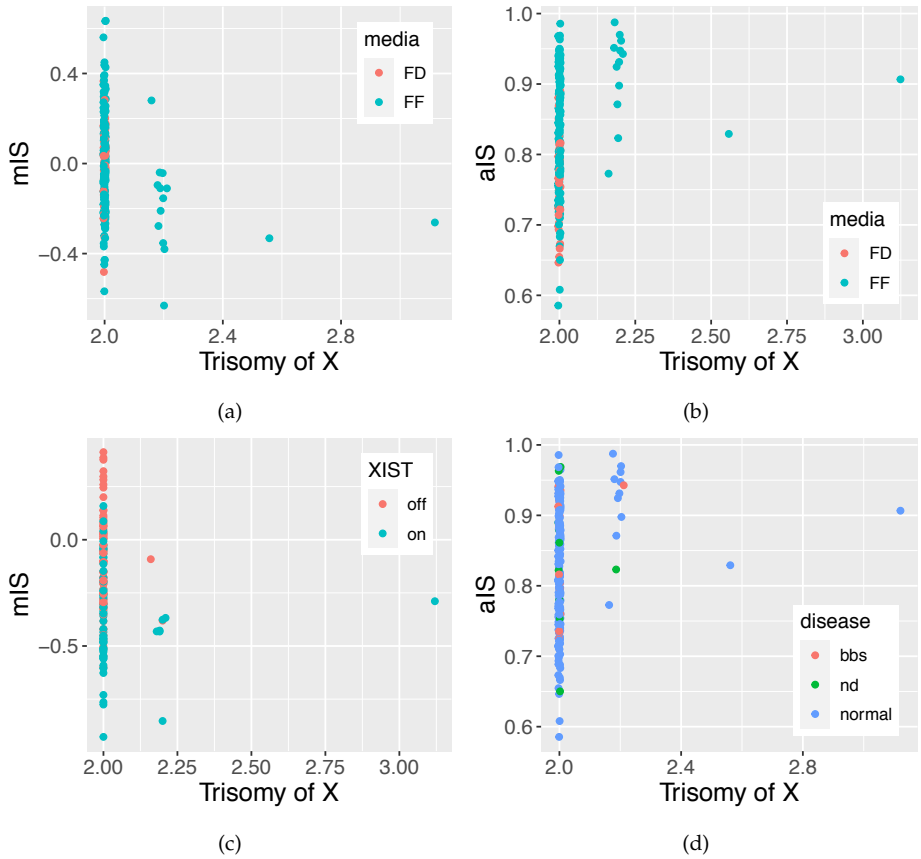


Figure 3.10: Trisomy of X chromosome occurs in 14 (5.1%) female h-iPSC lines in HipSci. a and b. The distribution of mIS (a) and aIS (b) with X chromosome trisomy. c. 12 out of 14 lines with X chromosome trisomy display on-mode XIST. Stratified by the mode of XIST, X chromosome trisomy is not associated with XCI level ($p > 0.2$ for both XCI metrics). d. The vast majority of lines (12 out of 14) with X chromosome trisomy were generated from healthy donors (bbs: Bardet-Biedl syndrome; nd: neonatal diabetes; normal: healthy)

3.4.2 Copy number alterations (CNAs) between h-iPSCs and fibroblasts

As discussed at the beginning of this section, sub-chromosomal CNAs were found in 41% of h-iPSCs generated from both genders, located on all chromosomes (Kilpinen et al. 2017). To study the association between XCI variation and CNAs, it is important to know the recurrent level of CNAs in females h-iPSCs, for the reason that the role of a common CNA is more useful in biological and disease-related research than a rare one. For this research aim, a segmentation of detected CNA regions was executed, with the workflow shown in figure 3.11.

This computational approach was executed on each chromosome, permitting that each CNA region was counted for its occurrence in all lines. Among 273 female h-iPSCs, 91 h-iPSCs carry 146 CNA segments (16 on X chromosome; 130 on autosomes). Even though Kilpinen et al. 2017 observed 22% CNAs which occurred in more than one line for 711 h-iPSCs, the recurrent level of CNAs is lower in the studied 273 female h-iPSCs: 24 (16.4%) CNA segments occur in more than one h-iPSC line, among which 3 (2%) segments occur in three lines. This difference might result from a smaller sample size and/or the single gender of lines in this thesis. The 24 recurrent CNA segments distribute on both X chromosome (7 segments) and autosomes (22 segments), where the autosome with most recurrent CNA segments are chromosome 7, 12 and 16 (4 segments on each). One of three most recurrent CNA segments takes place on X chromosome whereas the other two on autosomes (chromosome 12 and 16).

Figure 3.12 clearly shows that, among all detected CNAs, the majority (80%) is only detected in one h-iPSC line. The potential effect of these recurrent CNA segments in the variability of XCI level is conducted with variance component analysis (VCA), presented in section 3.5.

3.5 Variance component analysis (VCA) identifies donor and XIST as the most important sources of XCI variation

In previous sections, the general association between multiple sources (experimental, donor specific, XCI related and trisomy of X chromosome) and XCI level was presented. Among all factors, cell culture media and XIST expression level showed a clear association with XCI variation (figure 3.1 a and figure 3.5).

Different with single association test which examines the association between one single factor (i.e. cell culture media) and XCI level, the variance component analysis (VCA) is a method to assess the variance in a dependent variable which is associated with one or multiple independent variables (Bland et al. 1986). VCA is commonly based on mixed linear model, which allows both cat-

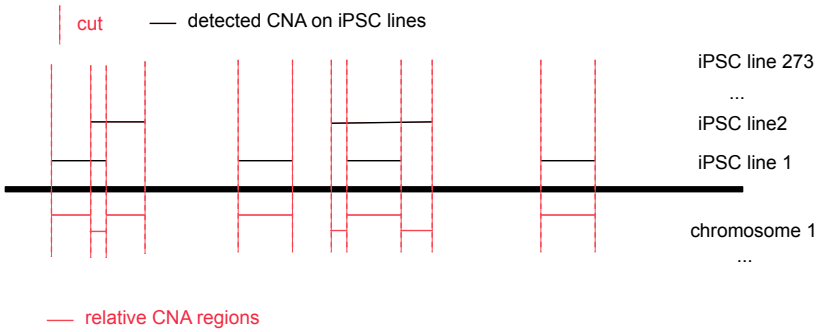


Figure 3.11: The process of CNA segmentation on one chromosome: order all detected CNAs by their start position and cut the chromosome region with the requirement that only one start position is allowed in one segmentation.

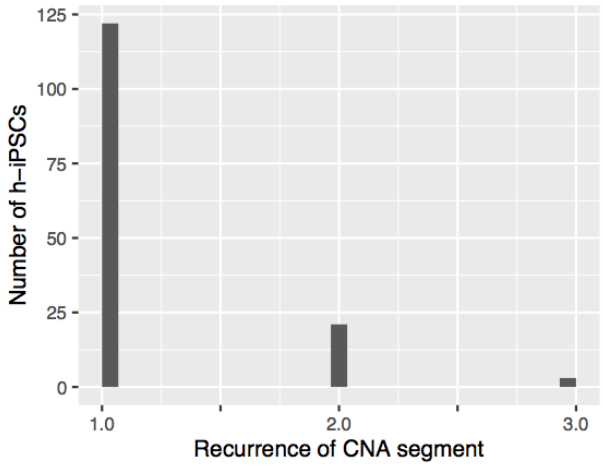
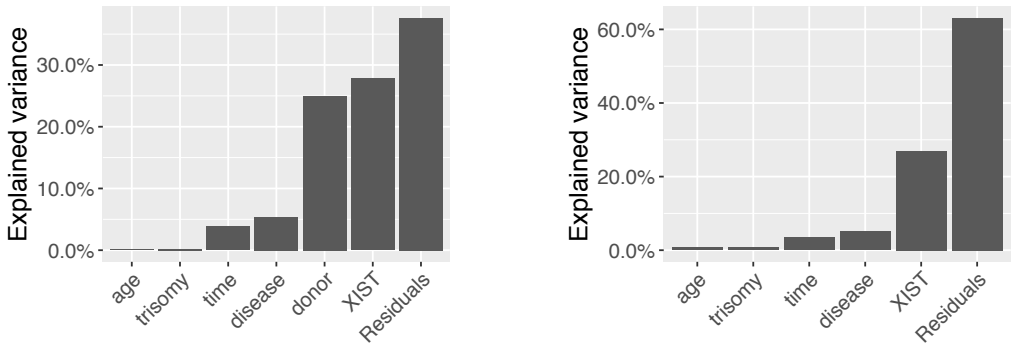


Figure 3.12: Overview of recurrent level of 146 CNA segments in 273 female h-iPSCs: 16.4% (24) CNAs occur in more than one h-iPSC lines, at the same time, 3 CNAs occur in three h-iPSCs, which is the maximum recurrent level.

egorical and continuous independent variables to be taken into account (X. Lin 1997, Kang et al. 2010). To limit the number of factors included in the mixed linear model, in this study, VCA was conducted in two steps: firstly, all factors presented in this chapter, except CNA segments, are fit in VCA model; secondly, factors which showed highest variance proportions in the first step, together with CNA segments were fit in VCA. The analysis was carried out with function *lmer* in package lme4 (D. Bates et al. 2014) and function *calcVarPart* in package variancePartition (Hoffman et al. 2016), with R version 3.4.0 (R Core Team 2017).

3.5.1 VCA without CNA segments

In the first step, VCA was carried out with mixed linear regression using following factors: cell culture time, age and health condition of donors, XIST expression level, as well as trisomy of X chromosome. To estimate the potential donor effect in h-iPSCs which were generated from the same donor, two VCA models were fit: the first one used all 273 female h-iPSCs and included the donor of h-iPSC as random effect; the second one used 205 female h-iPSCs which was randomly selected one line per donor thus the donor-effect was not included in the model.



(a) (b)

Figure 3.13: Variance component analysis (VCA) without CNA segments (Included factors: cell culture time, age and health condition of donor, XIST expression level, as well as trisomy situation of X chromosome. The corrected mIS value is used as XCI metrics. a. First model was fit for all 273 female h-iPSC lines, including donor as random effect. The first model identifies XIST expression ($>$ 25%) and donor (25%) as most important factors for XCI heterogeneity. b. Second model was fit for 205 randomly selected h-iPSCs with one line per donor. This model identifies XIST expression as the only important factor which explains 30% of variance in XCI level.

The first VCA model explains around 65% of variance of the XCI variation: among all tested factors, XIST (>25%) and donor effect (25%) are two factors

with highest proportion of variance explained (figure 3.13 a). In the second VCA model, XIST expression was identified as the only important factor (30% of variance), confirming the importance of XIST in XCI heterogeneity and the initial conclusion that it is essential to take the same-donor effect into consideration for researches using h-iPSCs (section 2.1.3). In these two VCA models, none of other tested factors shows a proportion of variance explained greater than 8%. As XIST expression and the same-donor effect showed their importance in the first level VCA, they were included in the second level VCA, where their contributions in the XCI heterogeneity were measured together with CNA segments.

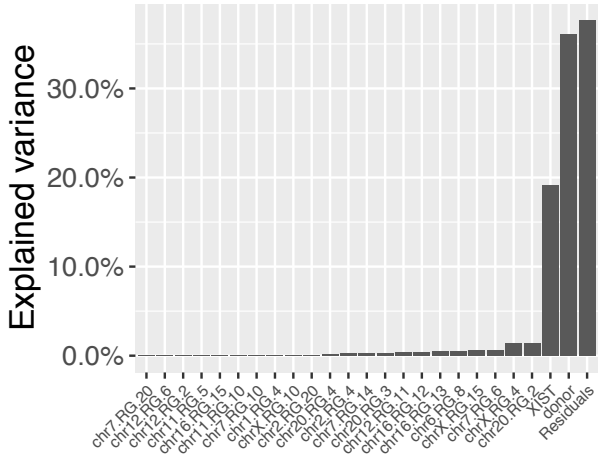
3.5.2 VCA with CNA segments

The second step of VCA used all 273 female h-iPSCs and included XIST expression, donor effect, as well as 29 recurrent CNA segments as independent variables. This model shows that none of CNA segment has a proportion of explained variance larger than 2.5%, meanwhile the donor effect counts for 35% of variance in XCI and XIST expression counts for slightly less than 20% of variance (figure 3.14 a). Another VCA model was exclusively fit with the three most recurrent CNA segments, XIST expression and donor effect, whereas all three CNA segments also showed proportion of explained variance smaller than 2.5% (figure 3.14 b).

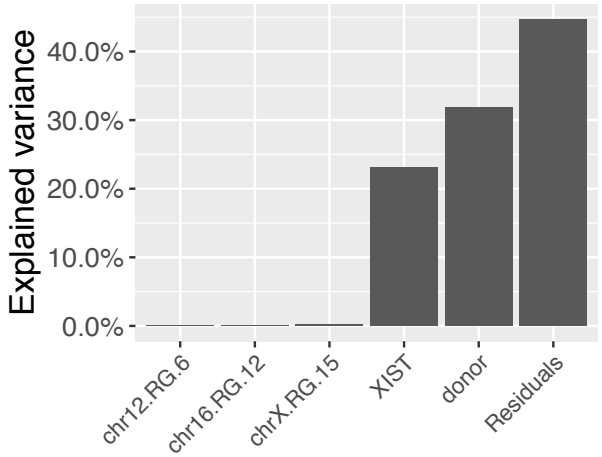
Apparently, as shown in figure 3.14, compared with CNA segments, even with the most recurrent ones, the donor effect and XIST expression level, which respectively explained 30% and 20% of XCI variation in two models, count for much more capacity in the variance explanation for XCI heterogeneity. The result reveals that, similar as trisomy of X chromosome, CNAs do not have a direct association with the XCI heterogeneity in female h-iPSCs. At the same time, the high proportion of explained variance by donor points to potential donor-specific determinants for the XCI level, which inspires my further analysis into the genetic variants.

3.6 Discussion: the expected and unexpected sources of XCI variation

After presenting the XCI heterogeneity in female h-iPSCs in chapter 2, to answer the question 'What causes the variation?' became my key concern. There has been a long time discussion about potential sources of XCI variation, mainly covering XIST, the culture time, the reprogramming process and the chromosomal variations (C. J. Brown, Hendrich, et al. 1992, Liang et al. 2013, Mekhoubad et al. 2012, Geens et al. 2016, Anguera et al. 2012, Briggs et al. 2015). Limitations of previous studies are clear: an insufficient number of samples and a lack of h-iPSCs from healthy donors as the reference. HipSci (Kilpinen et al. 2017) is



(a)



(b)

Figure 3.14: Variance Component Analysis used 273 female h-iPSCs and included XIST expression, donor effect and recurrent CNA segments. a. None of 29 recurrent CNA segments explains more then 2.5% of XCI heterogeneity. b. The 3 CNA segments with highest recurrent level also shows a proportion of explained variance less than 2.5% when they are included in VCA instead of all 29 CNA segments.

a great opportunity to investigate the sources of XCI variation, since all h-iPSC lines and all data were generated from the single institute, as well as that the cohort allows a big range of values for several factors (i.e. cell culture time and donor age). The result in this chapter can serve as the reference pool of causal factors for XCI variation in h-iPSCs.

XIST: not the perfect marker of XCI level

The gap between XIST expression level and the XCI status, is an unexpected result in this chapter. Particularly, when XIST expression is shut off, h-iPSC lines display a various XCI level, from proper XCI to complete XCI loss (section 3.3.2). Also, when taking h-iPSCs with high XIST expression into account, there are a certain number of lines with bi-allelic expression level, which is a sign of XCI loss (figure 3.5 b right bottom corner). The sharp drop of XIST expression during cell culture is a very interesting result. On one hand, it reveals that the mechanism of XIST might be time related, thus, there might be a time-dependent X chromosome activation mechanism. On the other hand, this result reproduces the result of previous studies which observed the XIST loss with small number of lines (Mekhoubad et al. 2012, Briggs et al. 2015, Anguera et al. 2012).

To summarize, even though there is a statistical association between XIST expression and XCI level, XIST is not a perfect marker for XCI, specifically for h-iPSCs in long cell culture.

The strong effect from cell culture media and the light effect from culture time

Surprisingly, the cell culture factor which has a strong effect on XCI level is the culture media, but not the long-discussed culture time. With HipSci, two independent analysis have shown the significant effect of cell culture media in h-iPSC lines. The first analysis shows the difference of pluripotency score by PluriTest (Müller et al. 2011) for h-iPSCs cultured in media FF and media FD, reported by Kilpinen et al. 2017. The second analysis shows the stratification of XCI level by different cell culture media, shown in section 3.1 of this chapter. These two results reveal that, the same culture media is necessary to maintain the homogeneity of h-iPSC lines. For scientists who are using h-iPSCs for their studies and who are going to reproduce a certain experiment with h-iPSCs, I recommend to keep the same culture condition.

Previous studies reported an erosion of XCI level in h-iPSCs with culture time (Mekhoubad et al. 2012, Anguera et al. 2012, Trokovic et al. 2015), however, this thesis had a different observation. When looking into XCI level of h-iPSCs in long culture time (up to 240 days), this thesis did not observe an erosion of XCI with time. Meanwhile, a slight XCI erosion was observed for relatively shorter cell culture (within 50 days, around passage 15 with cell culture FF, figure 3.2). Compared to this thesis, previous studies which observed XCI loss in culture included fewer number of h-iPSC lines (n = 11 in Trokovic et al. 2015 and n = 12

in Mekhoubad et al. 2012) and used shorted culture time (up to passage 21 and 24 in Trokovic et al. 2015 and Mekhoubad et al. 2012). In the larger set of 273 female h-iPSCs, the XCI level is not effected by cell culture time in long term, while at early passages, a slight loss of XCI with culture time is observed.

Genetic variations do not associate with XCI level

Female h-iPSCs with more than two copies of X chromosome display a various XCI level, from proper XCI to partial XCI loss (figure 3.10). This observation is consistent with the phenomenon that XCI process leaves one functional active X chromosome in cells regardless of existence of multiple X chromosomes (C. J. Brown, Ballabio, et al. 1991, Galupa et al. 2018). The other genetic variation, the sub-chromosomal alterations have a relatively low recurrent level in female h-iPSCs (only 20% occur in more than one line and the highest recurrent level is 3 lines) and do not show an association with XCI variation. Therefore, the genetic variations during the reprogramming of h-iPSCs are not the source of XCI heterogeneity at a general level.

No effect on XCI level from age and health condition of the donor

The donor metadata has been long-term missing in the study of XCI status in female h-iPSCs. Some studies with limited number of donors observed an erosion of XCI level in h-iPSCs generated from old female donors (Trokovic et al. 2015, Sardo et al. 2017). Here, with a range of donors from 5-year-old to 80-year-old, the XCI variation in h-iPSCs is found across all ages. This result reveals that the XCI loss is not gained or associated with age. Besides, the XCI variation exists in healthy donors of all ages, proving that the XCI heterogeneity is a general and common phenotype in h-iPSCs, thus the XCI level of h-iPSCs should be taken into account for further researches.

Inspiration: what is actually the donor effect?

VCA results show that there is a strong donor effect in the XCI heterogeneity (figure 3.13 a and figure 3.14). Then what is this donor effect? As shown in section 3.2, this donor effect is not age or health condition, then it might be genetics, which is the 'root' difference of human. Then I rephrase the previous question to: are genetic variants of the donor related to the XCI level in h-iPSCs?

This is an important question because it tries to find the connection between the genetic information on all chromosomes and the XCI, of which the understanding of mechanism is still limited. When preparing for this analysis, I read the paper by Luijk et al. 2018, which demonstrates the association between autosomal genetic variants and the methylation levels of CpG islands near XCI-escapees with large sample size ($n > 1,800$). This previous discovery, as well as results of this chapter inspired my investigation into the potential genetic determinants of the XCI variation, which is presented in chapter 4.

Chapter 4

Autosomal genetic determinants of XCI variation

- Any promising autosomal variants for the regulation of XCI?
Yes, meanwhile validation experiments are necessary.

In chapter 3, multiple factors were investigated for their association with XCI variability: major characteristics of donors (age and health condition); experimental preparation of h-iPSCs (cell culture time and media); XCI related gene expression level (XIST expression), as well as the genetic difference between h-iPSCs and fibroblast cells where they were generated (trisomy situation of X chromosome and copy number alterations on all chromosomes). In the analyses in chapter 3, a strong donor effect was found for the XCI heterogeneity in female h-iPSCs, meanwhile, this effect is not an 'outside' factor such as the age or the health condition of the donor. Inspired by these previous results, here I investigate whether autosomal genetics associate with the XCI variability.

Some studies present the association between autosomal genetic variants and the XCI level in females with sex-linked disease and in healthy females. For instance, Vianna et al. 2020 reports autosomal variants which affect XCI escapees in female intellectual disability patients with high level of XCI skewing ($> 90\%$) and Luijk et al. 2018 identifies autosomal loci associated with female specific X chromosome methylation with large cohort of males and females ($n > 1,000$ each group). In this chapter, instead of comparative analysis between proper XCI and highly skewed XCI, I execute regression models to study the contribution of genetic variants in the XCI variability.

Previous studies of genetic variants in h-iPSCs include Schwartzentruber et al. 2018 which studies the functions of variants in iPSC-derived neurons (123 h-iPSCs), Panopoulos et al. 2017 which presents disease-related variants based on family structures (222 h-iPSCs from 41 families), as well as DeBoever et al. 2017 which presents the effect of genetic variants on gene expression level using 215 h-iPSC lines in Panopoulos et al. 2017. This study is a primary discovery of XCI

related genetic variants in large-scale healthy and independent individuals.

The investigation of genetic determinants of XCI is conducted at two aspects: firstly, whether an autosomal genetic variant can directly effect XCI loss level; secondly, whether an autosomal genetic variant can regulate XCI loss via autosomal gene(s). The design of the first part is direct: for all 273 female h-iPSC lines in the study, to execute a genome wide association study (GWAS) between all autosomal variants and XCI metrics. In this GWAS model, two XCI metrics are used as phenotypes respectively. Compared to the first part of analysis, the second part is more 'targeted'. In Kilpinen et al. 2017, an expression quantitative loci (eQTL) analysis was carried out to identify variants which are associated with genes. The XCI variation is associated with gene expression at whole genome wide, with 85% of associated genes on autosomes (details in chapter 5). I subset variants which are associated with XCI-related genes and apply a linear model to test the association between each variant in this subset and XCI level. When a genetic variant is identified to be significantly associated with XCI level in this model, it might be possible to assume a causal path linking this genetic variant, its associated gene and the XCI level.

4.1 The 166 female h-iPSCs is a good representation of the female lines in HipSci

In chapter 1, I introduced that there were in total 273 female h-iPSCs from 205 independent donors enrolled in this project. In previous analysis, I have tried to make the use of most possible h-iPSC lines. Here, limited by the data availability, the analysis was executed on a subset of 166 female h-iPSCs in HipSci. Furthermore, all these 166 female h-iPSCs were enrolled in the eQTL analysis which identified the gene-variants association by Kilpinen et al. 2017. To check whether this subset is a good representation of the initial data set, I present the distribution of three XCI related factors, namely mIS, aIS and XIST, of these 166 female h-iPSCs in figure 4.1 (a, b and c).

The subset of 166 female h-iPSCs included in the GWAS analysis is a good representation of female h-iPSC population as lines with all three patterns of XCI have been included: proper XCI (approximate minimum mIS or maximum aIS), intermediate XCI loss level (intermediate value of mIS or aIS), or complete XCI loss (approximate maximum mIS or minimum aIS). Figure 4.1 shows that the distribution of mIS in these 166 females is similar to the distribution in 273 female h-iPSCs (figure 2.5 b). The association between mIS and aIS in this subset (Pearson correlation -0.47, figure 4.1 d) is similar as the association in 205 female h-iPSCs selected by one line per donor (Pearson correlation $= -0.5$).

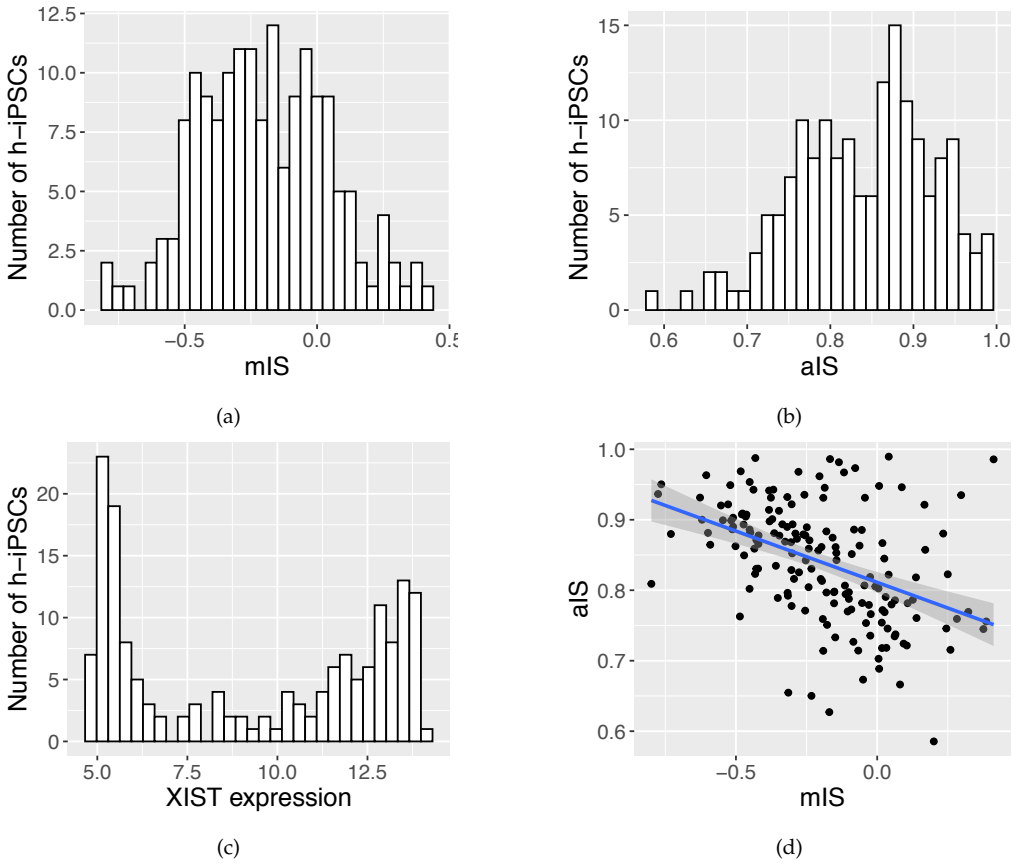


Figure 4.1: Summary of XCI level in 166 female h-iPSCs studied in this chapter. a. The distribution of mIS. b. The distribution of aIS. c. The distribution of XIST expression. d. The association between mIS and aIS (Pearson correlation $= -0.47$, p-value = 1.4×10^{-10}).

4.2 Use GWAS analysis to identify XCI associated autosomal variants at genome-wide

A genome-wide association study (GWAS) is the analysis of the association between genetic variants and a trait at population level. The trait in GWAS can refer to a disease or a phenotype in samples. In medical research, specifically in oncological research (Freedman et al. 2011, Pharoah et al. 2013), GWAS is widely used in the identification of disease related alleles.

Here, I apply GWAS to identify genetic variants which are associated with XCI variation with 166 female h-iPSCs from healthy and independent donors. The the number of samples involved in this analysis ($n = 166$) is limited, compared to other GWAS research of human phenotypes ($n > 3,000$ in Fadista et al. 2016, Tennessen et al. 2012 and Jian Yang et al. 2010). To reduce the bias introduced by the limited sample size, two XCI metrics were used respectively as the phe-

notype in GWAS and the overlap of top-associated variants with these two XCI metrics was extracted and interpreted. Considering that the effect of genetic variants on XCI level might be relatively small, two XCI metrics are used to validate the results. The expression level of XIST, the key factor of XCI process (Lyon 1961, Pontier et al. 2011, C. J. Brown, Hendrich, et al. 1992), is also investigated with GWAS as a XCI-related phenotype.

The GWAS analysis was conducted with the tool PLINK, version 1.90 (Purcell et al. 2007). The preparation of data input consisted of several steps: to convert data format from VCF files to PLINK format (.map and .ped file), to prune variants based on linkage disequilibrium (LD), to filter variants by their minor allele frequency (MAF) and to identify population structure by the principle component analysis (PCA) for the correction of population structure. For each step of preparation and the execution of GWAS, a Rmarkdown file together with a bash script were written and will be published with the manuscript of this thesis. These processes were executed on the HPC clusters of Wellcome Trust Sanger Institute.

4.2.1 Data preparation and the execution of GWAS

The preparation of input files followed guidelines of PLINK (PLINK 2010). Vcftools (version 0.1.17, Danecek, Auton, et al. 2011) was also used during preparation process to transform VCF files of exome sequencing data to format required by PLINK (.map and .ped files).

Preparation of phenotype data

As required by the PLINK manual (PLINK 2010), the phenotype file consists the basic information of h-iPSCs (the family ID, the individual ID and the gender of lines) and the three XCI related phenotypes. Since all 166 female h-iPSCs in GWAS were from independent donors, the family ID was set the same as individual ID, meaning that family was not a confounding factor in GWAS. Meanwhile, all 166 h-iPSCs were generated from females, so the sex information was ignored in the phenotype file, using the function *allow–no–sex* in the execution of PLINK.

Preparation of genotype data

The genotype of the h-iPSCs in this project were analysed by an Illumina HumanCoreExome-12 BeadChip. The experimental process, as well as the genotype calling and imputation process are presented in chapter 1 (section 1.1.1), summarized from Kilpinen et al. 2017.

The original VCF file of genotypes, which contains 8,600,656 variants, was transformed to PLINK required format, using the function *plink* in vcftools (Danecek, Auton, et al. 2011). The function *recodeA* was called to recode genotype matrix for all 166 female h-iPSCs, where genotypes of each variant were recoded to 0, 1 or 2 according to the reference allele, meanwhile this reference allele was added to the end of variant ID. For instance, rs1857-61220_G stands for the variant rs185761220 with allele G as the reference, whereas the genotype 0, 1 or 2 stands for the number of allele G in the h-iPSC line. In this chapter, this format of variant ID is used for the presentation of results.

Filtering of variants

Filtering of variants was based on the minor allele frequency (MAF) of variants in the data set of 166 female h-iPSCs. According to the definition by the National Center for Biotechnology Information (NCBI, https://www.ncbi.nlm.nih.gov), MAF is the frequency at which the second most common allele occurs in a given population.

Common values of MAF threshold in previous population-based genetic researches are 0.01 (i.e. Cassa et al. 2013), 0.05 (i.e. HapMap consium, Consortium et al. 2005) and 0.1 (i.e. the study of human height by Jian Yang et al. 2010). For this project, the MAF threshold was set to 0.05, which means that variants of which the MAF is smaller than 0.05 were removed. After filtering, 6,449,949 variants were remained.

Pruning variants based on linkage disequilibrium

In population genetics, linkage disequilibrium (LD) refers to the non-random association of alleles of different loci (Slatkin 2008). LD can occur on neighbouring loci due to physical connection, as well as on loci on different chromosomes, since it is influenced by multiple factors, for instance genetic recombination, the mutation rate and the system of mating (Lewontin et al. 1960, Hill et al. 1968). The existence of LD brings statistical problems in GWAS: single-nucleotide polymorphisms (SNPs) in strong LD tent to have similar p-values in the association study which brings difficulty to the identification of true causal variant of the phenotype or disease (Korte et al. 2013). Different computational approaches have been developed to address this problem (Chapman et al. 2003, Balding 2006), whereas in PLINK the variance inflation factor (VIF) is used to identify and to remove correlated variants.

By the definition of James et al. 2013, VIF is the quotient of the variance in a model with multiple terms by the variance of a model with one term alone. VIF measures the collinearity where multiple variables are correlated in the regression model and contain similar information of variance (Dormann et al. 2013). In genetics, VIF is used to measure how independent SNPs are from one another, with formula 4.1 in the manual of PLINK (PLINK 2010):

$$\frac{1}{1-R^2}, \tag{4.1}$$

where R^2 is the multiple correlation coefficient for a SNP being regressed on all other SNPs simultaneously.

For a certain SNP, when its VIF value equals to 1, which implies $R^2 = 0$, this SNP is completely independent of all other SNPs; when its VIF value is greatly larger than 1, for instance, equals to 10 which implies $R^2 = 0.9$, this SNP is correlated with other SNPs and has collinearity problems in the standard multiple regression analysis. The PLINK manual (PLINK 2010) recommends the use of a VIF threshold value between 1.5 and 2, to remove correlated SNPs and to maintain enough number of variants. Here, VIF threshold is set to 2 for pruning variants.

The function *indep* of PLINK pruned variants by recursively removing SNPs within a sliding window of which VIF values are greater than the VIF threshold. Parameters in the pruning process were set as follow: the window size in SNPs as 50, the number of SNPs to shift the window at each step as 5 and the VIF threshold as 2. After the filtering process and the pruning process, 1,241,616 variants were kept for further process from the original 8,600,656 variants.

Principle component analysis of 166 female h-iPSCs

As described in (Kilpinen et al. 2017), donors of h-iPSCs were research volunteers recruited from the National Institute for Health Research (NIHR) Cambridge BioResource. In the cohort, the vast majority of recruited volunteers are from the UK, while a small proportion are from other continents (i.e. Africa and Asia).

The previous study with the same h-iPSCs by Kilpinen et al. 2017 used population structure as the random effect factor in the research of genetics. Here, to investigate whether the population structure needs to be adjusted, the principle component analysis (PCA) was carried out for 166 female h-iPSCs using pruned and filtered variants, with the *pca* function in PLINK. The default setting of the *pca* function extracts the top 20 principal components (PCs) of the variance-standardized relationship matrix. Results of PCA include eigenvectors which are written to .eigenvec file, and eigenvalues which are written to .eigenval file. The PCA result suggest that the vast majority of h-iPSCs (163 out of 166) can be grouped in one cluster, while the remaining in the second cluster (figure B.3). Therefore, the first component was used as a covariate in GWAS to adjust for the population structure in 166 female h-iPSCs.

Execution of GWAS analysis

GWAS analysis between processed variants and each of three phenotypes was conducted with linear regression (function *linear* in PLINK). The linear regression contained PC1 of PCA results as a covariate to adjust for the population structure in the data set. The *adjust* function of PLINK was applied to have results after multiple testing with six correction methods: Bonferroni (Bonferroni 1936, Dunn 1961), Holm step-down (Holm 1979), Sidak single-step (Šidák 1967), Sidak step-down (Ludbrook 1998), Benjamini-Hochberg (**BHcorrect**), as well as Benjamini and Yekutieli (Benjamini and Yekutieli 2001). To keep the consistency in the whole analysis, results with Benjamini-Hochberg correction are used for the interpretation.

4.2.2 Results of GWAS: no significant variant for single phenotype but overlap top-associated variants across phenotypes

Results of GWAS are checked at three levels: firstly, an overview of result for each of the tested phenotypes, including most XCI-associated variants and the distribution of these variants on chromosomes; secondly, the overlapping of 100 most associated variants between proxy mIS and aIS; thirdly, the variance component analysis with identified XCI-associated variants to test their capacity to explain the variance of XCI.

Firstly, I investigate whether any variants are significantly associated with tested phenotypes after multiple testing control (FDR $<$ 10%) or with raw p-value smaller than 10^{-8}, which is the most commonly accepted threshold in human genome analysis (C. Xu et al. 2014). Compared to previous human genome studies (Fadista et al. 2016 with 12,590 individuals, Jian Yang et al. 2010 with 3,925 individuals and Tennessen et al. 2012 with 15,585 individuals), the sample size in this project is limited (166 individuals). Considering the bias that this small sample size brings to the statistical test, I also look into the overlap of 100 variants with the smallest raw p-values in GWAS with two XCI metrics, as the overlap variant(s) may also be informative for the effect of genetics on the XCI level.

The reason to check the overlap between mIS and aIS, but not XIST is that mIS and aIS are both XCI metrics which summarize the XCI status of h-iPSCs from different aspects (mIS: methylation; aIS: expression) and that they have shown a good association between themselves (Pearson correlation = -0.5 in the data set of 273 female h-iPSCs, figure 2.9 a). Even though high expression of XIST (on mode) refers to proper XCI level in theory (Pontier et al. 2011), the correlation between either mIS or aIS and XIST expression for h-iPSCs with on mode XIST is weaker than the correlation between proxies (Pearson correlation between mIS and on-mode XIST: -0.24; Pearson correlation between aIS and on-mode XIST: 0.31, chapter 3). To avoid the potential bias, the GWAS result with XIST is checked separately to identify variants related to 'on/off mode' of XIST.

GWAS result for XCI metric mIS

After multiple testing correction with Benjamini-Hochberg method (Benjamini and Hochberg 1995), there was no variant significantly (FDR < 10%) associated with mIS. The smallest p-value before multiple testing correction was 5.86×10^{-7} (variant rs185761220, chromosome 4). Among the top six variants associated with mIS, four of them had raw p-value smaller than 10^{-6}. These six variants locate on five chromosomes: two variants on chromosome 4 (rs185761220, rs60320061) and one variant on each of chromosome 3 (rs12632135), chromosome 14 (rs34518442), chromosome 16 (rs11248915) and chromosome 17 (rs9904875). Figure 4.2 shows the distribution of mIS on different genotypes of these six variants and figure 4.3 is the Manhattan plot to present the significance of genetic variants in this GWAS.

Even though a differential distribution of mIS is observed on these variants, specifically on variant rs185761220 (chromosome 4, raw p-value in GWAS = 3.4×10^{-7}) and on variant rs12632135 (chromosome 3, raw p-value in GWAS = 4.7×10^{-6}), no evidence is found in previous studies which can support the role of these variants in the XCI process in h-iPSCs. In section 4.2.3, variance component analysis (VCA) is applied to estimate the contribution of these variants to the XCI variability.

GWAS result for XCI metric aIS

Similar to the GWAS result with mIS, no variant show significant association (FDR $<$ 10%) with aIS after multiple testing control (Benjamini-Hochberg method). Before adjustment, all top six variants had raw p-value in GWAS at 10^{-6}. These six variants also locate on different chromosomes: one variant on chromosome 1 (rs357207, p-value = 1.1×10^{-6}), one variant on chromosome 9 (rs75781423, p-value = 2.3×10^{-6}), one variant on chromosome 11 (rs112901333, p-value = 3.1×10^{-6}) and three variants on chromosome 2 (rs79068464, p-value = 3.4×10^{-6}; rs145753116, p-value = 3.8×10^{-6}; rs199934696, p-value = 4.3×10^{-6}). The distribution of aIS on these variants is shown in figure 4.5 and the significance level of all variants in this GWAS is shown in figure 4.4.

Similar to the section 4.2.2, no previous studies reported the role of these six variants in the XCI process. The contributions of these six variants to the XCI variability are estimated by VCA in section 4.2.3.

The overlap between most XCI-related variants

There are four variants overlapping between 100 variants most associated with either mIS or aIS in GWAS analysis (ranked by raw p-value in GWAS), which are rs79084431, rs79131540, rs75734556 and rs74554399. These four variants all locate on chromosome 1, at nearby positions (from 218144672 to 218182410). The distribution of two XCI proxies on these four variants is shown in figure 4.6.

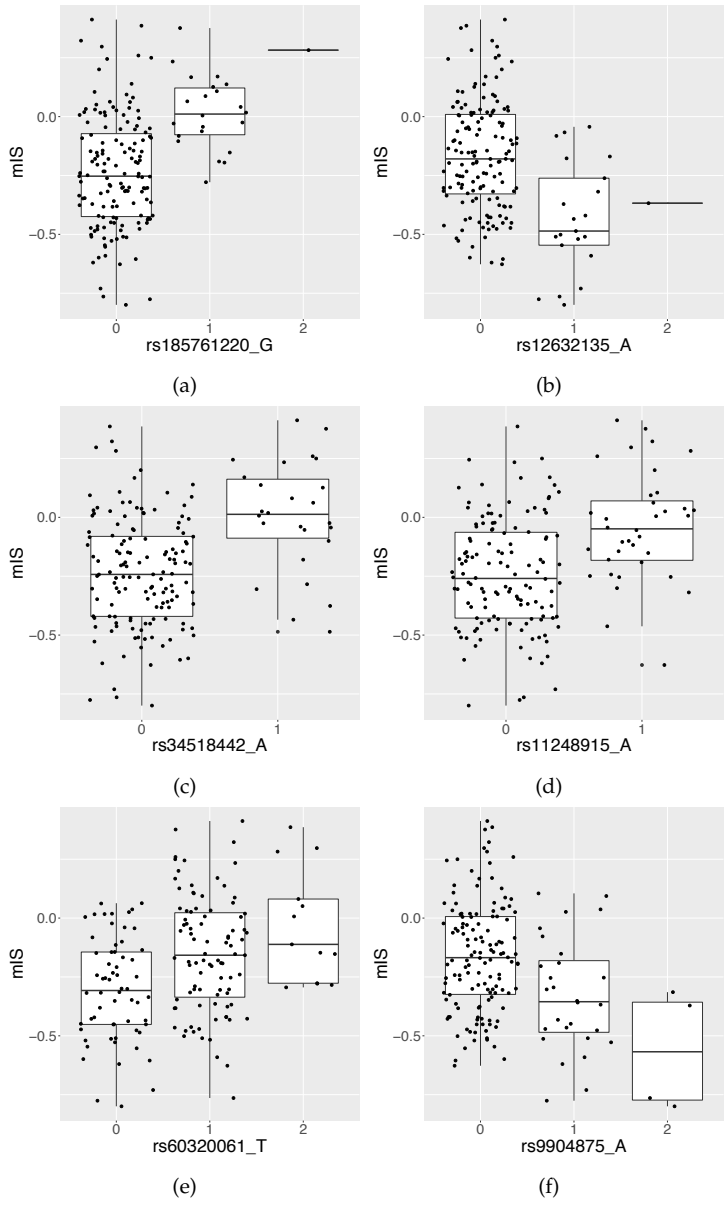


Figure 4.2: Distribution of mIS on genotypes of six variants which are most associated with mIS in GWAS using 166 female h-iPSCs. a. rs185761220, locates on chromosome 4 (raw p-value in GWAS = 3.4×10^{-7}). b. rs12632135, locates on chromosome 3 (raw p-value in GWAS = 4.7×10^{-6}). c. rs34518442, locates on chromosome 14 (raw p-value in GWAS = 2.4×10^{-6}). d. rs11248915, locates on chromosome 16 (raw p-value in GWAS = 6.0×10^{-6}). e. rs60320061, locates on chromosome 4 (raw p-value in GWAS1.1×10^{-5}). f. rs9904875, locates on chromosome 17 (raw p-value in GWAS1.1×10^{-5}).

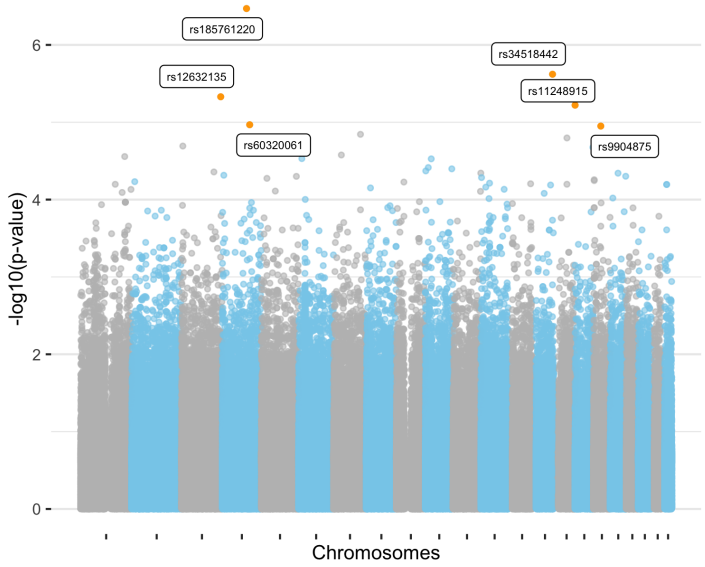


Figure 4.3: Manhattan plot showing the GWAS result when mIS is used as XCI metric. The six variants with highest p-values in GWAS are colored in orange and annotated.

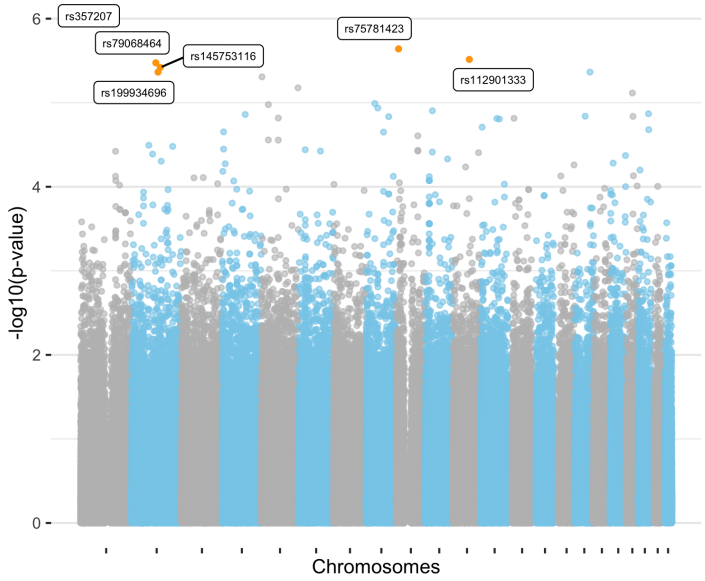


Figure 4.4: Manhattan plot showing the GWAS result when aIS is used as XCI metric. The six variants with highest p-values in GWAS are colored in orange and annotated.

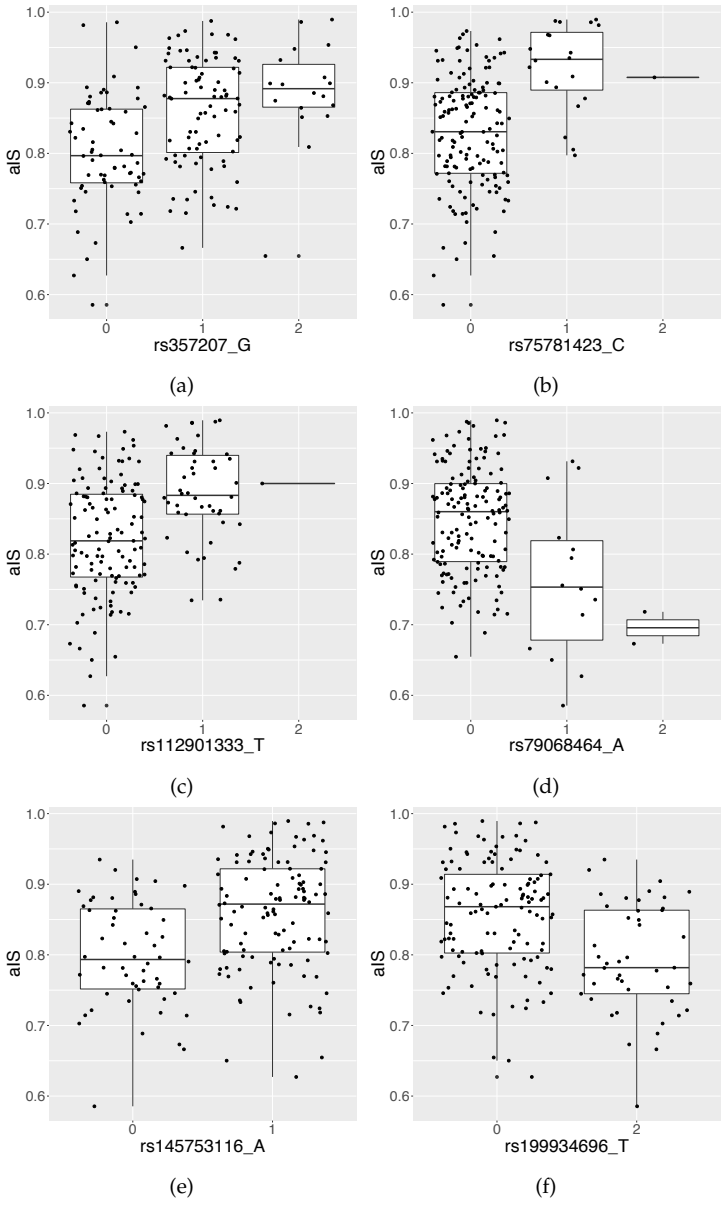


Figure 4.5: The distribution of aIS on genotypes of six variants which are most associated with aIS in GWAS using 166 female h-iPSCs. a. rs357207, locates on chromosome 1 (raw p-value in GWAS = 1.1×10^{-6}). b. rs75781423, locates on chromosome 9 (raw p-value in GWAS = 2.3×10^{-6}). c. rs112901333, locates on chromosome 11 (raw p-value in GWAS = 3.1×10^{-6}). d. rs79068464, locates on chromosome 2 (raw p-value in GWAS = 3.4×10^{-6}). e. rs145753116, locates on chromosome 2 (raw p-value in GWAS = 3.8×10^{-6}). f. rs199934696, locates on chromosome 2 (raw p-value in GWAS = 4.3×10^{-6}).

According to ENSEMBL human genome reference, version GRCH37/hg19 (Harrow et al. 2012), all four variants are located very close to gene AL355526.1 (ENSG00000230714), which is a long non-coding RNA (LncRNA). With version GRCH38 (Harrow et al. 2012), gene AL355526.1 is mapped to a larger region, which includes all these four variants. The close location of these four variants also shows its effect on the distribution of XCI proxies: the distribution of either mIS or aIS is very similar on these four variants (figure 4.6). Based on this observation, these four variants are considered as one variant region which is associated with XCI level.

Both ENSEMBL human genome website (www.ensembl.org) and GeneCards (www.genecard.org) show that gene AL355526.1 is a long non-coding RNA, meanwhile neither of these two websites or other literature presents the association between gene AL355526.1 and the XCI process or X chromosome related phenotypes. Similar with variants most associated with XCI metrics, the identified overlapping variant region is included in VCA models (section 4.2.3) to test its capacity to explain XCI variability.

GWAS result for XIST

After multiple testing control (Benjamini-Hochberg method, Benjamini and Hochberg 1995), five variants are found significantly associated with XIST expression (FDR $< 10\%$).
Four out of these five variants have raw p-value smaller than 10^{-8}. Locations and corrected p-values of these five variants are listed as following: rs145753116 and rs199934696 (chromosome 2, 5×10^{-6} and 5.4×10^{-5} , respectively), rs2009-5981 (chromosome 16, 5.4×10^{-5}), rs117389731 (chromosome 19, 1×10^{-4}) and rs143-848756 (chromosome 12, 2×10^{-2}). The association between these five variants and XIST expression is shown in figure 4.7.

Previous results present that XIST expression drops sharply at day 50 in culture (chapter 3,figure 3.8 a). Therefore, I color h-iPSCs in previous result by culture time (short: $<$ 50 days; long: $\geqslant$ 50 days), shown in figure 4.8. Unfortunately, for all these five variants, the vast majority of h-iPSCs which have genotypes associated with low XIST expression level was cultured in long culture time. Therefore, considering the time effect on XIST (section 3.3.3), it is impossible to draw the conclusion about the effect of these five variants on XIST expression.

4.2.3 Biological conclusion: estimation of genetic effects in XCI variation with VCA

In Chapter 3, I used variance component analysis (VCA) to study the proportion of XCI variance explained by different non-genetic factors. Here, VCA models are used to analyse the contribution of genetic variants to XCI variability. The

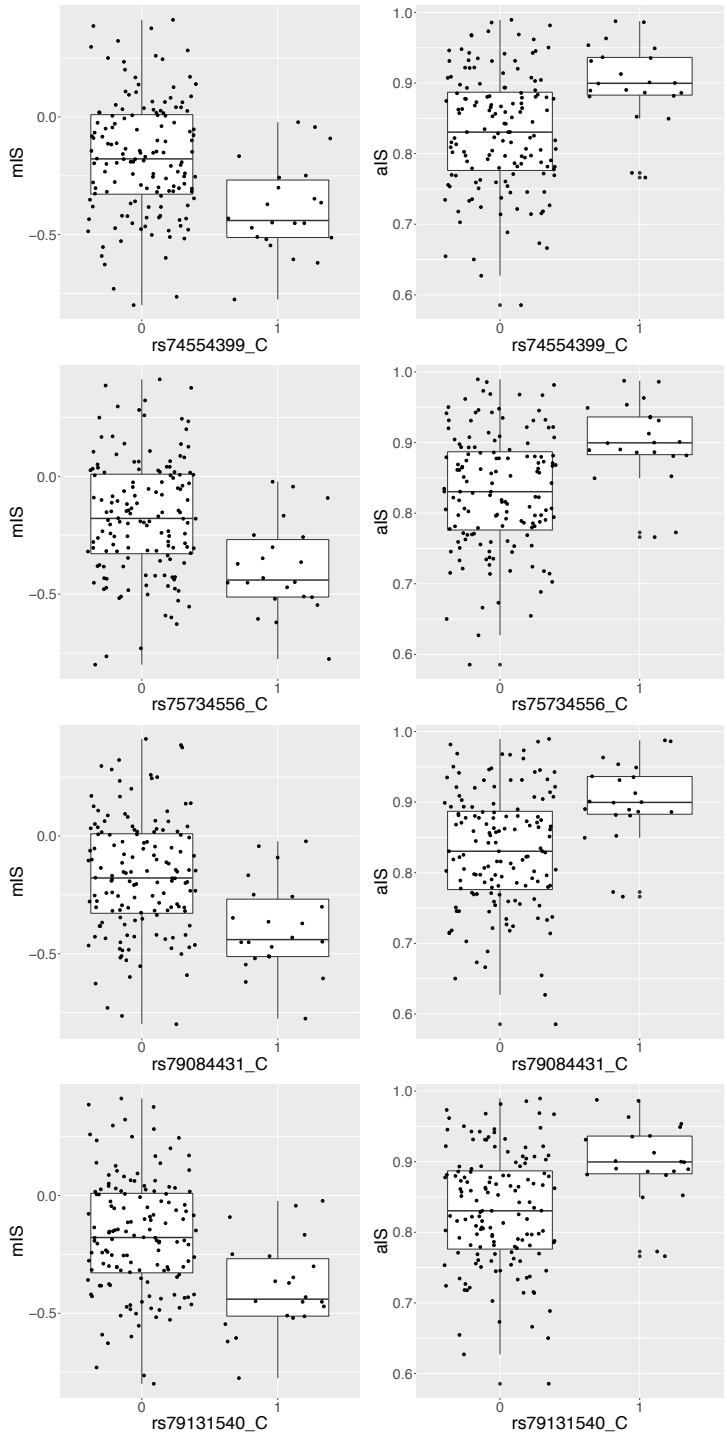


Figure 4.6: The distribution of mIS and aIS on four overlapping variants between XCI-related 100 variants with smallest raw p-values in GWAS analysis (left: mIS; right: aIS), showing a consistent direction of effect of variants on two XCI metrics.

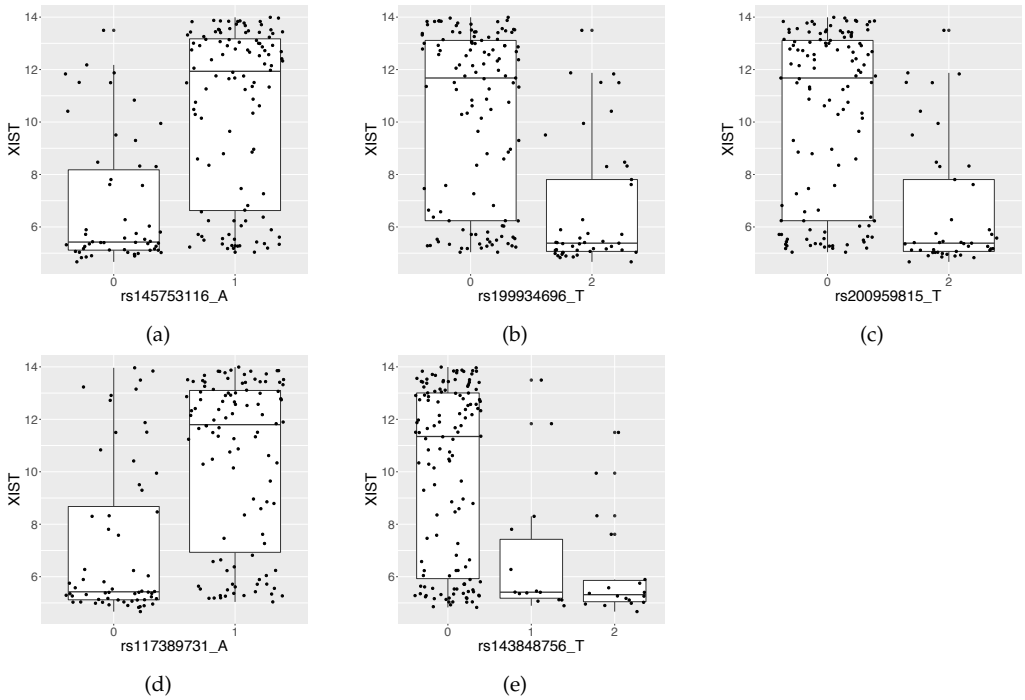


Figure 4.7: Five variants are significantly (FDR < 10%) associated with XIST expression in GWAS using 166 female h-iPSCs. a. rs145753116, locates on chromosome 2 (corrected p-value = 5×10^{-6}). b. rs199934696, locates on chromosome 2 (corrected p-value = 5.4×10^{-5}). c. rs20095981, locates on chromosome 16 (corrected p-value = 5.4×10^{-5}). d. rs117389731, locates on chromosome 19 (corrected p-value = 1×10^{-4}). e. rs143848756, locates on chromosome 12 (corrected p-value = 2×10^{-2}).

investigation is carried out in two steps: in the first step, to study the six variants which are most associated with either of two XCI metrics; in the second step, to study the overlapping variant region between XCI metrics. According to results in chapter 3, all studied factors can explain approximately in total 60% of variance of XCI. Among all studied factors, the donor effect and XIST expression were the only two factors which can explain more than 20% of the variance of XCI loss. As 166 female h-iPSCs are all from independent donors, the donor effect was not included in the VCA model in this section. Therefore, in this section, VCA models are applied to estimate the effect of genetic variants and XIST expression in the variance of XCI.

Before the analysis with genetic variants, I fit VCA models with only XIST expression for both two XCI metrics to study whether the XIST expression would maintain the same capacity in variance explanation as the result in chapter 3, where XIST expression counted for 27% of variance of mIS in the model with 273 female h-iPSCs and approximately 30% variance of mIS in the model with 205 female h-iPSCs (without donor effect). With 166 female h-iPSCs, the propor-

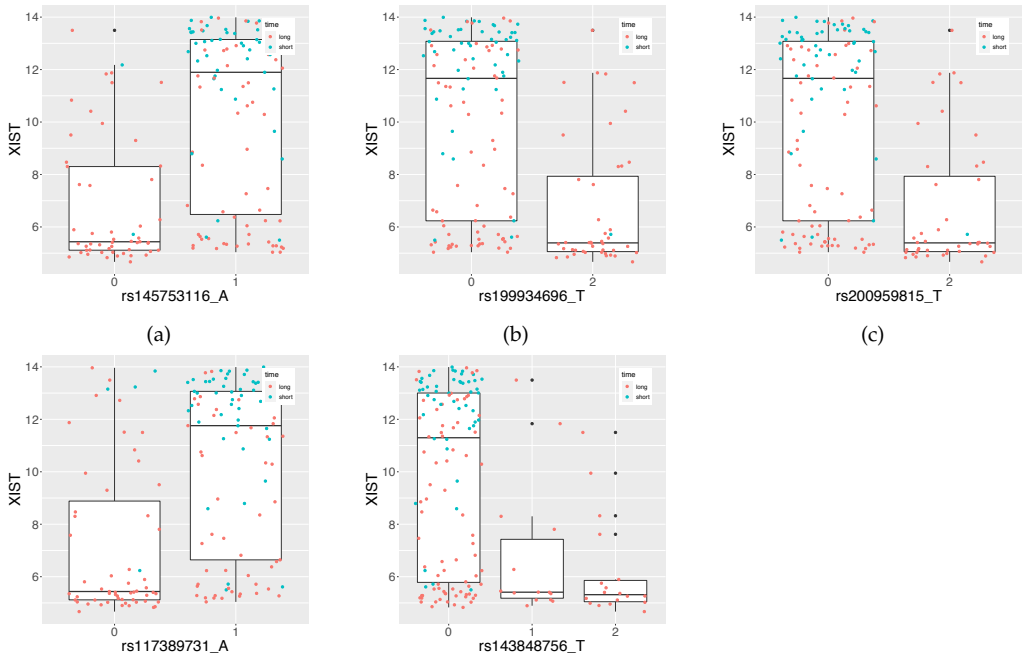
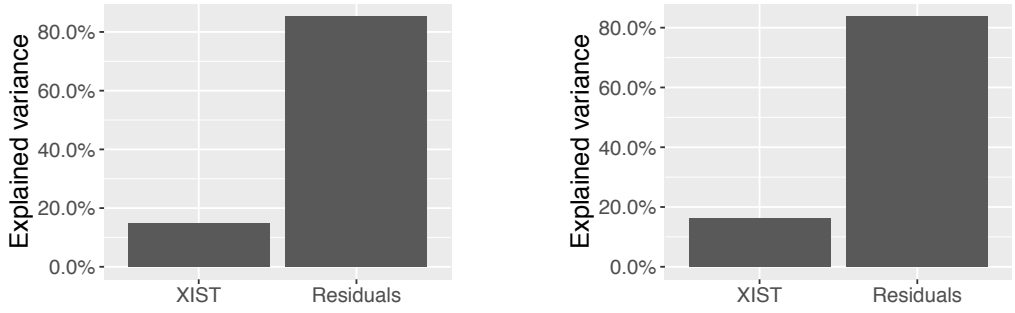


Figure 4.8: The distribution of XIST on five variants which are found significantly associated with XIST in GWAS using 166 female h-iPSCs, with each h-iPSC colored by the cell culture time. The short culture refers to lines which were cultured shorter than 50 days (cyan) and the long culture refers to lines which were cultured more than 50 days (red).

tion of variance explained by XIST is reduced: XIST expression count for 15% of variance in mIS and 16% of variance in aIS (figure 4.9).

This reduction is taken into account for the interpretation of further VCA models, specially for the comparison of different factors for their capacity to explain the XCI variation. The detailed design of the VCA model is as follow: firstly the VCA model was fit respectively for each of two XCI metrics, namely mIS and aIS, with XIST expression and top six genetic variants which were most associated with this metrics in GWAS; secondly, another VCA model was fit for each XCI metric using XIST expression and top six variants with either XCI metrics. As there was no overlap between top six variants for two XCI metrics, the total number of genetic variants involved in the second analysis was twelve. Results of these two designs of models are shown in figure 4.10 (a and b: first VCA models; c and d: second VCA models).

The first-part VCA results show that, for a certain XCI metrics, top XCI-related genetic variants identified by GWAS are able to explain a good proportion of its variance: in total, all six variants explained approximately 62% of variance for mIS and 49% of variance for aIS, meanwhile the most-associated variant ex-



(a) mIS (b) aIS

Figure 4.9: The result of VCA for 166 female h-iPSCs from independent donors with only XIST expression. a. Using mIS as XCI metrics (XIST counts for 15% of variance). b. Using aIS as XCI metrics (XIST counts for 16% of variance).

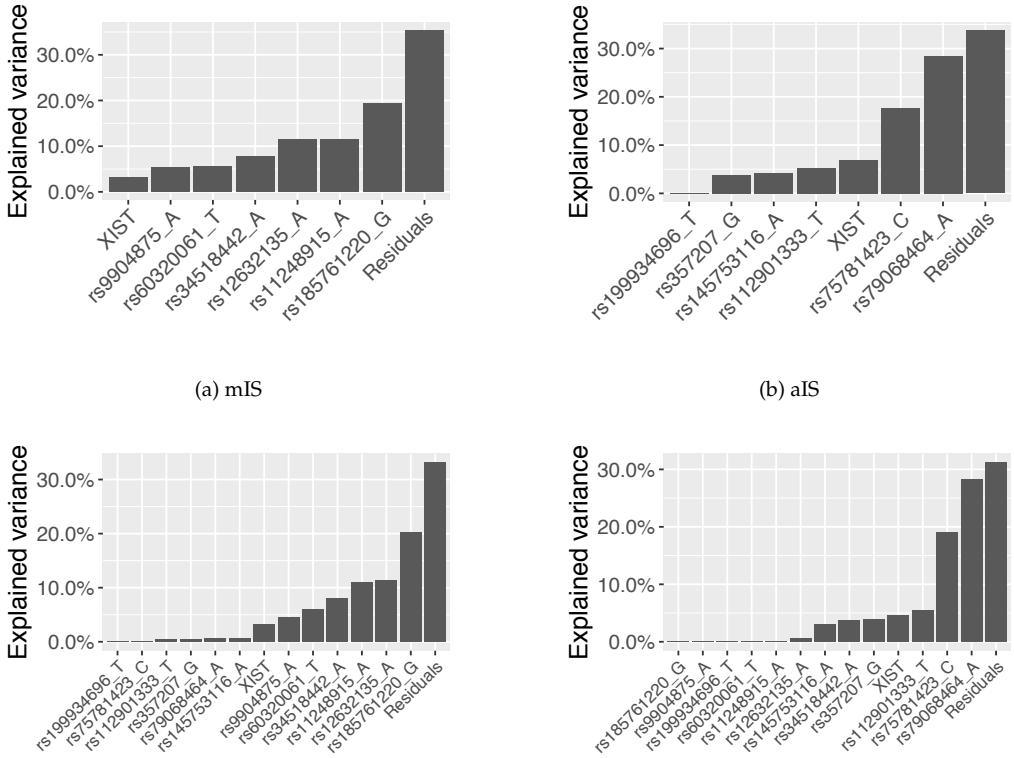
plained approximately 20% of variance for mIS and slightly more than 30% of variance for aIS (figure 4.10 a and b).

When comparing the proportion of variance explained by each variant and by XIST expression in the VCA model, all six variants associated with mIS and four out of six variants associated with aIS counted for higher proportion than XIST expression, which was the key factor to explain XCI in chapter 3. As shown in the VCA results, the impact of genetic variants on XCI variation was surprisingly high, even if the effect by XIST was taken into account.

In the first-part VCA model, the genetic variants which explained the highest proportion of XCI variance were rs185761220_G (approximately 20%) and rs79068464_A (31%), respectively for mIS and aIS. The difference of proportion explained by six variants in the VCA model for mIS is much smaller than the difference in the model for aIS, where the proportion explained by variant rs79068464_A is bigger than the sum of the proportion explained by all other factors (approximately 21%).

Regardless of the high proportion of explained variance by genetic variants in the first-part result, genetic variants failed to explain the variance of the other XCI-metrics in the second-part VCA since no variant counted for more than 1% of XCI-variance, shown in figure 4.10 (c and d). The failure of cross-explanation of genetic variants reveals a gap for the study of genetic effects in XCI variation: a true and significant XCI determinant should show its capability to explain the XCI variation in both two XCI metrics, like donor effect and XIST in VCA results in section 3.5.

To investigate whether there is any variant which explains a good proportion of both two XCI metrics, another VCA model includes the overlap variant re-



(a) mIS

(b) aIS

(c) mIS

(d) aIS

Figure 4.10: Unignorable effects from genetic variants for the XCI variation detected by VCA in 166 female h-iPSCs. a. VCA for mIS using XIST and top 6 genetic variants associated with mIS in GWAS; b. VCA for aIS using XIST and top 6 genetic variants associated with aIS in GWAS; c. VCA for mIS using XIST and top 6 variants associated with mIS and top 6 variants associated with aIS in GWAS; d. VCA for aIS using XIST and top 6 variants associated with aIS and top 6 variants associated with aIS in GWAS

gion, which refers to the four variants which are overlapped between top 100 variants associated with mIS and with aIS. As discussed in section 4.2.2 (figure 4.6), the four overlapping variants are considered as one genetic region given that they are physically located next to each other and the distribution of mIS or aIS is the same on them. The new VCA model includes top 6 variants associated with single XCI metrics and the overlap variant region.

In the new model, the capability to explain the XCI variation of the overlap region is quite different for mIS and aIS: for mIS, the overlap region counts for around 2% of variance which is the least among all tested factors (figure 4.11 a); while for aIS, this region counts for more than 5% which is slightly higher than the proportion explained by XIST expression level (figure 4.11 b).

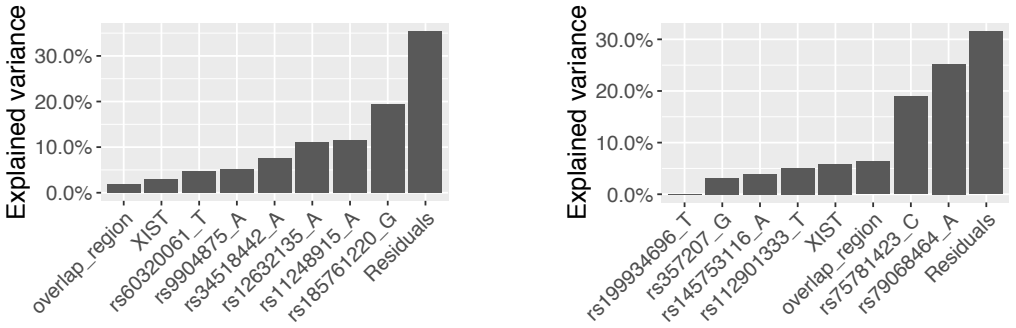
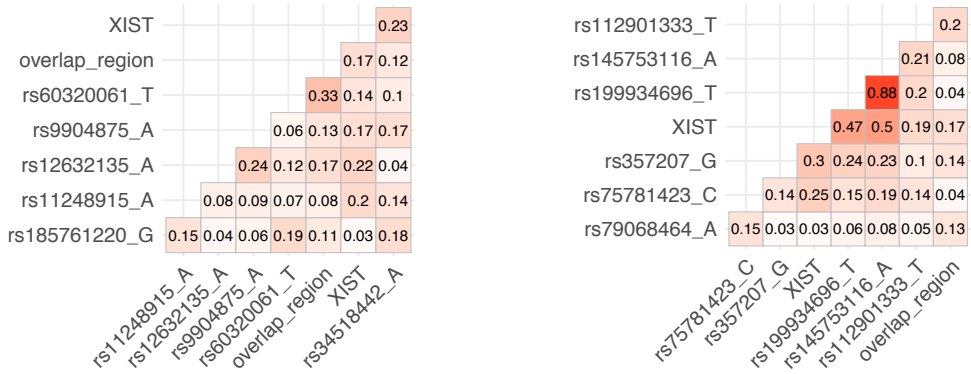


Figure 4.11: The investigation of whether the overlap variant region is a common source of variance for two XCI metrics using VCA model. a. The overlap region counts for 2% of variance of mIS. b. The overlap region counts for slightly more than 5% of variance of aIS.

To study the potential reason behind this variance, I checked the correlation between the overlap variant region and the twelve variants which were most associated with two XCI metrics (figure 4.12). The overlap region has slight correlation with two single variants, namely rs60320061 (Pearson correlation = 0.33) and rs112901333 (Pearson correlation = 0.20), respectively associated with mIS and aIS (figure 4.10). Meanwhile, XIST correlate with two aIS-related variants (rs199934696 and rs145753116, Pearson correlation > 0.4), which might confound the VCA result (figure 4.12 b). Therefore, the VCA result for mIS would be more trustful, which reveals a relatively small proportion of explained variance by the overlap variant region.

To summarize, VCA models show that these genetic variants count for unignorable proportion of XCI variation (in total 62% for mIS and 49% for aIS), at the same time there is a failure of cross-explanation of variants for the other XCI metrics (less than 5% explained by variants associated with the other XCI metrics). When the overlap variant region is included in the VCA model, it shows a small capacity to explain the mIS (2%).



(a) mIS (b) aIS

Figure 4.12: Correlation matrix between factors in VCA models including XIST expression, the overlapping region and six genetic variants most associated with XCI proxies for mIS (a) and aIS (b).

The validation is always important for the genetics research. The common validation is the proof by other researches and the validation experiment. For genetic variants which are found most associated with XCI metrics by GWAS and the overlap variant region between the top 100 variants identified by two GWAS, no previous studies reported their effect in XCI process or potential role in X-linked diseases. Among studied variants, three variants are most valuable for the validation experiment: rs185761220 (most associated variant with mIS), rs79068464 (most associated variant with aIS) and the overlap variant region. The validation of the role of these variants can be executed with CRISPR-CAS9 technology (Doudna et al. 2014, Ran et al. 2013, G. Wang et al. 2017) which knocks off single or multiple variants. To achieve a reasonable statistical power, the validation experiment requires a sufficient number of h-iPSC lines, as well as a solid CRISPR-CAS9 application experiences in h-iPSC lines. An example of sample size estimation is given in section 4.4.

Due to a limitation of these resources, the experimental validation work is not done in this thesis, nevertheless, I am looking forward to further researches to study the role of these genetic variants in the XCI process and to reveal the biological mechanism behind the association between autosomes and the X chromosome. Results in this section also reveal the difference between methylation-based XCI metric and expression-based metric, which is also observed in chapter 2. Since DNA methylation is part of the mechanism of the X-silencing (Moore et al. 2013, Phillips et al. 2008), I expected a close association between these two XCI metrics. However, according to studies of this thesis, an inconsistency between the methylation level and the expression level is observed, showing as the biased correlation between metrics, as well as the different associated autosomal genetic variants. For the XCI in h-iPSCs, there is still no consensus about the regulation of gene expression by the XCI-related methylation. I expect to see more

biological researches to explore the mechanism and to answer this question with h-iPSC lines.

4.3 The association test between XCI and important variants detected by eQTL

At the beginning of this chapter, I presented that there are two ways to identify genetic determinants of XCI variation: the first one is to apply GWAS to look for variants which are associated with XCI proxies at the whole genome level (section 4.2); the second one is to firstly have a subset of variants who are indirectly associated with XCI variation and then test the association between these variants and the XCI level.

This section presents the process and the result of the second method, which makes use of the list of XCI-associated genes (chapter 5) and the eQTL result from Kilpinen et al. 2017. The eQTL analysis in Kilpinen et al. 2017 summarized autosomal variants which are associated with autosomal genes for 239 h-iPSCs (166 donors, RNA-sequencing data). After correction of donor effect and of multiple testing with Benjamini-Hochberg method (Benjamini and Hochberg 1995), 4,347 variants are found associated with 4,422 genes (Kilpinen et al. 2017). These variants locate on all autosomes: most on chromosome 1 (426 variants) while least genes on chromosome 21 (70 variants). The eQTL analysis reveals a direct association between autosomal genetic variants and autosomal gene expression levels. Since previous studies found DNA variants have an important role in the regulation of gene expression level (Pai et al. 2015, Cheung et al. 2009), the eQTL analysis in Kilpinen et al. 2017 reveals regulator-variants for these 4,422 genes.

With the work of this thesis, the XCI variation is found to be associated with the expression of 1,757 autosomal genes using XCI metric mIS and 2,013 autosomal genes using XCI metric aIS (FDR < 10%, details in chapter 5). In this section, I extract the regulator-variants ($n = 272$ for mIS and $n = 388$ for aIS) of these XCI-related genes from the eQTL result by Kilpinen et al. 2017 and study the association between XCI level and this subset of genetic variants.

4.3.1 Extraction of genetic variants associated with XCI-related autosomal genes

The association analysis between gene expression and XCI level is presented in chapter 5. Briefly, genes whose standard deviation across 273 h-iPSCs was smaller than 0.1 were removed, leaving 41,353 genes for further analysis. Among these 41,353 genes, 39,856 autosome genes were included in the association analysis with XCI proxies. With linear regression between expression level of each gene and two XCI metrics (mIS and aIS), after multiple testing control (Dabney et al. 2010, FDR < 10%), the number of autosomal genes associated with XCI level is 1,757 for mIS and 2,013 for aIS.

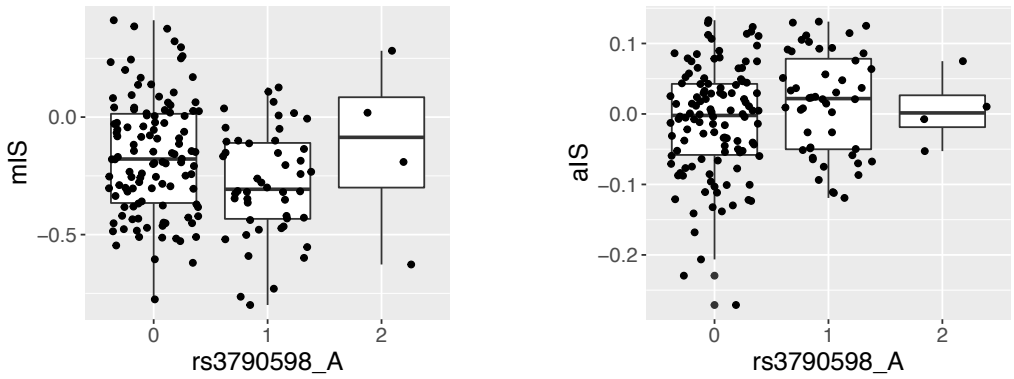
Among these genes, 261 mIS-related genes and 376 aIS-related genes were found associated with autosomal variants (272 and 388, respectively) in the eQTL analysis by Kilpinen et al. 2017 (FDR < 10%). Unfortunately, six variants which were found most associated with either of XCI metrics and five variants which are found associated with XIST in section 4.2 are not significantly associated with autosomal genes, thus not involved in this analysis.

The genotype matrix of these variants were extracted from plink-transfered files of 166 female h-iPSCs (section 4.2.1). As written before, the genotype of each variant was recoded to 0, 1 or 2 according to reference allele. Moreover, the reference allele was added to the end of variant ID. In the rest of this chapter, I use this format of variant ID for discussion.

4.3.2 Association analysis between genetic variants and XCI proxies

The association analysis between genetic variants and XCI metrics was done with univariate linear regression. After multiple testing correction (Dabney et al. 2010), there is no variant significantly associated with either mIS or aIS: the smallest q-value in the analysis is 0.49 for mIS and 0.32 for aIS.

Regardless of non-significant result, there are 22 variants overlapping among 100 variants which are most associated with the two XCI metrics. The variant rs3790598, which has the smallest q-value among these 22 variants, shows a good probability to have causal effect on XCI (figure 4.13).



(a) (b)

Figure 4.13: Distribution of mIS (a) and aIS (b) on different genotypes of variant rs3790598.

This variant is located on chromosome 1 and is found associated with RNA helicase gene MOV10 (figure 4.14 a) according to eQTL analysis by Kilpinen et al. 2017. For variant rs3790598, 114 out of 166 female h-iPSCs have genotype 0, 48 h-iPSCs have genotype 1 while only 4 h-iPSCs have genotype 2. As the number

of h-iPSCs with genotype 2 on variant rs3790598 is very limited, my study is focused on the association between XCI level and the other two genotypes.

Figure 4.13 shows that the association between genotypes of the variant rs3790598 and XCI level is consistent between two XCI metrics: h-iPSCs with genotype 1 tend to have proper XCI, which appears as low mIS value and high aIS value, respectively standing for higher methylation level and higher mono allelic expression level.

The expression of MOV10 is associated with variant rs3790598 with q-value equals to 4.9×10^{-5} in the eQTL result by Kilpinen et al. 2017, referring to 2.1×10^{-7} in simple ANOVA test: h-iPSCs with genotype 0 of variant rs3790598 had higher expression level of MOV10 (figure 4.14 a). Meanwhile, high expression level of MOV10 is related to proper XCI, showing as low mIS (Pearson correlation = -0.25) and high aIS (Pearson correlation = 0.21) in figure 4.14 (c and d). Since h-iPSCs with genotype 0 of variant rs3790598, which is the genotype with higher MOV10 expression level, show a higher mIS value, there is a conflict between variant-XCI association and gene-XCI association. To investigate what might be confounding factor in this association, I plot the distribution of XIST on genotypes of variant rs3790598 where a higher XIST expression level is found in h-iPSCs with genotype 1 (p-value $= 0.09$, ANOVA test, figure 4.14 b).

The association between MOV10 and XIST was observed by P. J. Kenny et al. 2014 using human embryonic kidney cells, where the knock-down of MOV10 by immunoprecipitation (IP) resulted in an enrichment of XIST, meanwhile the MOV10-binding regions on XIST were observed by the individual nucleotide resolution Cross-Linking and ImmunoPrecipitation (iCLIP). Combining this information and analyses in this section, I suppose that the regulation process of genetic variant rs3790598 on XCI is complex, which might be a combination of regulation via both MOV10 and XIST. To have a more detailed study of this causal path of XCI, more data sources are necessary.

4.4 Discussion: a potential causal relation between autosomal variants and XCI

In this chapter, I use statistical methods to discover potential genetic determinants of XCI variation in female h-iPSCs. Using genome wide association study (GWAS), six variants which are most associated with either of two XCI proxies are found promising, regardless of the non-significance level (FDR < 10%, Benjamini-Hochberg method, Benjamini and Hochberg 1995). With variance component analysis (VCA), I find that the variant rs185761220 which is most associated with mIS in GWAS and the variant rs79068464 which is most associated with aIS in GWAS explain the most portion of XCI variability (approximately 20% and slightly more than 30%, respectively). These two variants are capable to explain the XCI variability slightly more than XIST expression, which

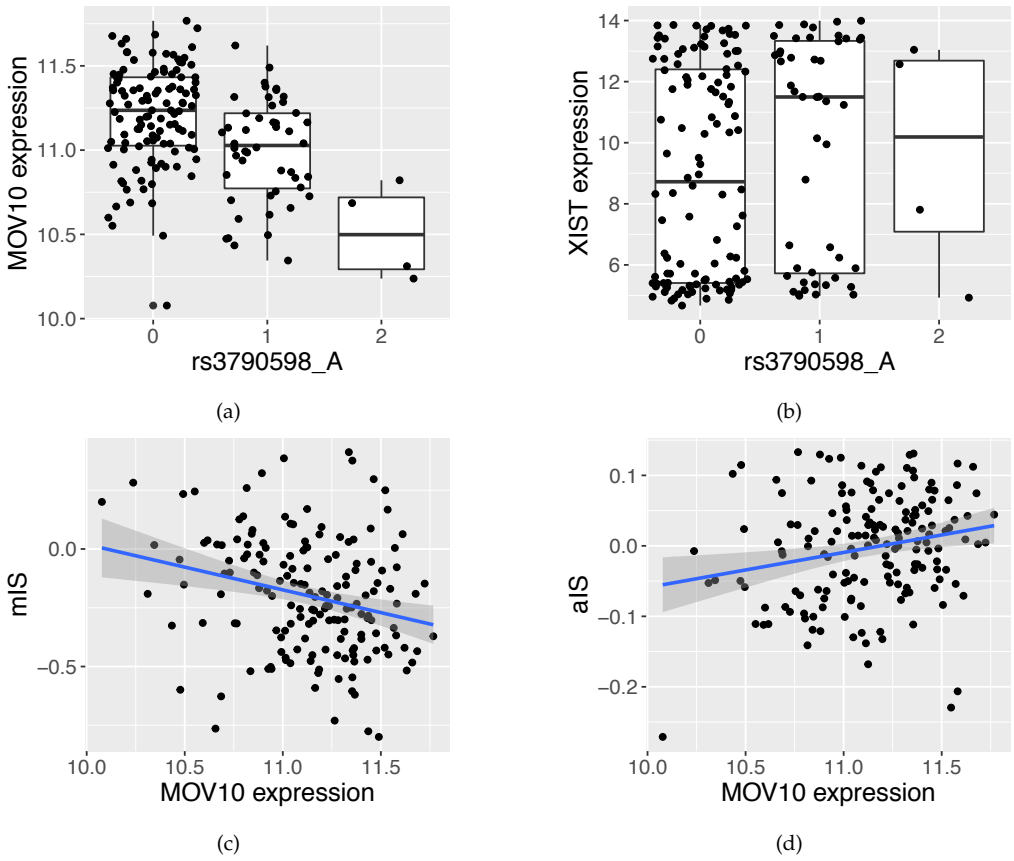


Figure 4.14: Mechanism behind the association between genetic variant rs3790598 and XCI loss. a. H-iPSCs with genotype 0 has higher expression level of MOV10 (q-value = 4.9×10^{-5}, eQTL by Kilpinen et al. 2017). b. H-iPSCs with genotype 1 has higher expression level of XIST (p-value = 0.09, ANOVA test). c. The association between MOV10 expression level and mIS (Pearson correlation = -0.25, p-value = 0.001). d. The association between MOV10 expression level and aIS (Pearson correlation = 0.21, p-value = 0.05).

was the most important factor in VCA models of chapter 3.

Using expression quantitative trait loci (eQTL) result by Kilpinen et al. 2017, I extract autosomal genetic variants are associated with XCI-related genes (FDR < 10%). With univariate linear regression, I identify the potential causal path connecting variant rs3790598, gene MOV10 and XCI level. MOV10 is an RNA helicase and is found to associate with XIST expression in human embryonic kidney stells (P. J. Kenny et al. 2014). Therefore, the regulation of XCI via this causal path would be an interesting topic.

Since this thesis is the first population-level study of XCI status in h-iPSCs from the single data source (HipSci), it is difficult to validate results here with previous studies. To prove the causal or the regulative effect of these genetic variants, biological experiment would be necessary. One of the most widely used technologies in similar type of researches would be the knock-off of variants using the CRISPR-CAS9 technology (G. Wang et al. 2017, Ran et al. 2013, Doudna et al. 2014). The potential experimental validation could be the knocking-off of these variants in h-iPSCs and the measurement of altered XCI loss level afterwards. For such experiments, I estimate the minimum number of h-iPSCs to be included to ensure a proper statistical power. Below is an example of the estimation for the potential causal path.

This potential causal path is between variant rs3790598, gene MOV10 and XCI level. To test this path, the variant rs3790598 needs to be knocked-off for h-iPSCs whose genotype is 1 (one allele with A). The XCI level and expression level of MOV10 need to be measured for these h-iPSCs before and after the knock-off. In the previous analysis with all 166 female h-iPSCs, the average of aIS value for h-iPSCs with genotype 0 is -0.01, while the average of aIS for h-iPSCs with genotype 1 is 0.01. To observe such a reduction in aIS in the experiment, with significance level $\alpha = 0.05$, statistical power $1 - \beta = 0.8$, a minimum of 32 h-iPSCs is needed. The average of MOV10 expression level is 11.0 in h-iPSCs with genotype 1 and 11.2 in h-iPSCs with genotype 0, with the same statistical requirement, a minimum of 17 h-iPSCs is needed for the experiment.

Either 32 or even 17 is a large number of lines required for experiments using h-iPSCs, because the setting up of experiment is time consuming and because either methylation array or RNA-sequencing needs to be done for all tested lines before and after knock-off. The long lasting and large scale data measurement would make the validation experiment an entire project. For this reason, the experimental validation of these variants is not executed in the work related to this thesis. For the paper which I am going to publish for this project, I am seeking for collaborations to carry out knock-down experiment of gene MOV10 on three or four h-iPSCs to observe the alteration in XCI level. Furthermore, I expect to see more experimental researches on the effect of autosomal variants on the XCI process.

Chapter 5

Consequences of XCI heterogeneity in female h-iPSCs

- What is the direct consequence of XCI variation?
The alteration of gene expression.
- Can XCI by inherited by h-iPSC derived cells?
Yes.

Chapter 2 presented XCI heterogeneity using 273 female h-iPSCs from HipSci (Kilpinen et al. 2017): 1% (four) lines display a complete loss of XCI while other lines have different XCI level, showing with variation in both methylation level and bi-allelic expression level of the X chromosome. In h-iPSCs and in other mammalian cells, XCI is the dosage compensation mechanism to balance the sex-related genes in two genders, specifically to ensure the similar expression level of X-located genes in males and females (Lyon 1961, Brockdorff et al. 2015, Heard et al. 1997, Avner et al. 2001, Galupa et al. 2018). Therefore, regarding the XCI heterogeneity in female h-iPSCs (chapter 2), two questions arise: does this XCI variability have functional consequences in gene expression level and to what extent are these consequences?

The overexpression of X-linked genes following the loss of XCI was observed in h-ES cells by Bar et al. 2019 and in h-iPSCs by Brenes et al. 2020, which also reported an increase level of protein but not messenger RNA (mRNA) for autosomal genes. Inspired by these studies, this chapter studies the association between XCI variation and genome wide expression alteration, especially the difference of alteration pattern for X-linked genes and for autosomal genes.

H-iPSCs have the unique ability to differentiate to human somatic cells (Hu et al. 2010b), which makes h-iPSCs and iPSC-derived cells very important tools in regenerative medicine as well as in scientific research (S. M. Wu et al. 2011, Knoepfler 2009, Singh et al. 2015, Castagné et al. 2011). Considering the variance of XCI in female h-iPSCs (chapter 2), it is very important to clarify whether the XCI heterogeneity in h-iPSCs can be inherited by iPSC-derived cells and if

yes, whether the XCI heterogeneity is associated with downstream features, for instance, the immune-related expression alteration in iPSC-derived cells. This chapter analyses 273 female h-iPSCs in HipSci (Kilpinen et al. 2017) and 43 iPSC-derived macrophages by Alasoo et al. 2018, mainly using the expression level to investigate the above questions.

5.1 XCI heterogeneity results in genome wide expression alteration

The RNA-sequencing data of 273 female h-iPSCs (Kilpinen et al. 2017) is used for the association analysis between the XCI variation and gene expression alteration. To remove genes which have constant expression level in the population, genes with standard deviation smaller than 0.1 are excluded, remaining 41,353 genes (39,856 autosomal genes and 1,497 X-linked genes) in the analysis.

The association test between gene expression and XCI level is done with a univariate linear model, using mIS as XCI metric, with formula 5.1:

$$\text{mIS} = \alpha + \beta\, \text{gene expression}_i, \tag{5.1}$$

where α and β stand for the intercept of the model and the coefficient of gene expression, i stands for gene_i to test in the model, with $i \in [1 : 41,353]$.

For the genome wide association study, 41,353 independent models were fit for mIS, with p-value of each model used as the significance level of association between the tested gene and mIS. After multiple testing correction (Dabney et al. 2010), the XCI loss level is found associated with 2,086 genes (FDR < 0.1), of which 329 (15%) genes are on the X chromosome and 1,757 (85%) genes are on autosomes, shown in figure 5.1. This result reveals that XCI variation leads to genome-wide expression alteration and this effect is not limited to genes on the X chromosome, but also genes on autosomes.

5.2 The XCI loss results in up-regulation of X-linked genes and random alteration of autosomal genes

After observed that genes on both X chromosome and autosomes have significant association with XCI loss (FDR < 0.1, 2,086 genes), the pattern of expression alteration is studied, using the β value from the linear model 5.1.

Consistent with the mechanism behind XCI (C. J. Brown, Hendrich, et al. 1992, Heard et al. 1997, Avner et al. 2001, Galupa et al. 2018) and with previous study by Brenes et al. 2020, I find that the XCI loss results in a different pattern of expression alteration in the X chromosome and in autosomes: on the X chromosome, most XCI-associated genes are up-regulated (97%, 320 genes), while

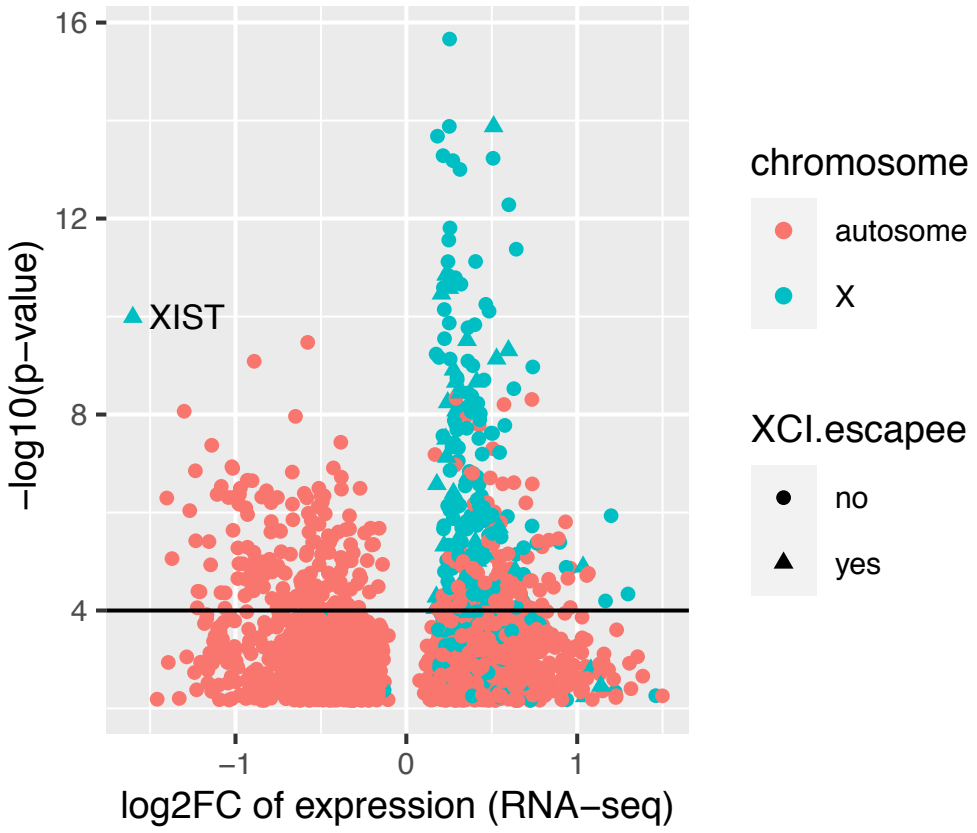


Figure 5.1: The genome wide gene alteration with XCI variation: 2,086 genes are XCI-related, including 329 X-related genes (15%, cyan) and 1,757 autosomal genes (85%, red). Each dot presents a gene included in the analysis, while x-value of the dot is the $\log_2$FC value of this gene and y-value of the dot is the $-\log_{10}$ of p-value of this gene in the univariate linear model (formula 5.1). $\log_2$FC is defined as the $\log_2$ transformed ratio of the maximum expression level of one gene over the minimum expression level of this gene in all female h-iPSCs and converted to positive value for up-regulated genes with XCI loss, to negative value for down-regulated genes. Genes with larger absolute $\log_2$FC value exhibit bigger range of expression in female h-iPSCs. The horizontal line stands for p-value = 10^{-4}. The known XCI-escapees are labeled by the shape of point (triangle: known XCI-escapee; round: not known XCI-escapee, details in section 5.2).

on autosomes, fractions of up- or down-regulated genes are almost equal (45% up-regulated). As introduced in section 1.2.3, Carrel et al. 2005 found that up to 25% of human X-linked genes can escape from the XCI process, namely XCI-escapees, while 15% of them showed a constant capability to escape from XCI across samples. Furthermore, Tukiainen et al. 2017 presented a systematic survey of XCI on 29 types of tissues from 446 individuals, where the incomplete XCI was found to affect the expression level around 23% of X-chromosome genes.

To incorporate known XCI-escapees in my study of gene regulation in female h-iPSC lines, 200 genes are taken from the study by Tukiainen et al. 2017, which were identified as XCI-escapees in at least one tissue and were checked for their overlap with XCI associated genes identified in 273 female h-iPSCs.

Figure 5.1 shows the regulation of genes with XCI-escapees labeled, where it is observed that XCI-escapees are associated with XCI loss at various levels of significance.

Among 329 X-linked genes which are found associated with XCI (FDR < 0.1, section 5.1), 22% (72 genes) are known XCI-escapees. After the removal of these known XCI-escapees, 257 X-linked genes are associated with XCI. In addition, 98% (251 out of 257) of X-linked genes are up-regulated (figure 5.1). Therefore, I present that the XCI loss in female h-iPSCs results in genome wide alteration of gene expression level. In addition, on the X chromosome the associated genes are mostly up-regulated and on autosomes there is an equal rate of up- and down-regulation.

5.3 Inherited XCI level in cells derived from h-iPSCs and its immune-related effects

The inheritance of expression and genetic signatures from h-iPSCs to iPSC-derived cells is a key concern for the use of h-iPSCs in disease modeling and cell therapies (Tiscornia et al. 2011, Doss et al. 2019, Singh et al. 2015, D'Antonio-Chronowska et al. 2019). The XCI of h-iPSC is an important X-related signature, thus whether the XCI level can be stably inherited in iPSC-derived cells is an important question for the research of h-iPSCs.

HipSci is a great opportunity to study the genetic similarities between h-iPSCs and iPSC-derived cells: Kilpinen et al. 2017 reported the genetic characteristics of 711 h-iPSCs and Alasoo et al. 2018 differentiated macrophages from 86 h-iPSC lines to study their genetic features and molecular functions. These two data sets enable the investigation of the inheritance stability of XCI level from h-iPSCs to iPSC-derived cells with a reasonable sample size. Furthermore, using iPSC-derived macrophages, it is possible to estimate the effect of XCI variation on the immune-related expression alteration.

5.3.1 XCI heterogeneity is stably inherited in iPSC-derived cells

The RNA-sequencing data are available for 43 over lapping h-iPSCs between Alasoo et al. 2018 and 273 female h-iPSCs studied in this thesis. The rIS is computed for iPSC-derived macrophages with the definition in section 2.3.2. Figure 5.2 presents the clear association between rIS of iPSC-derived macrophages and of their progenitor h-iPSCs, indicating a stable inheritance of XCI level by iPSC-derived cells (Pearson correlation $= 0.61$, p-value $= 1.3 \times 10^{-5}$).

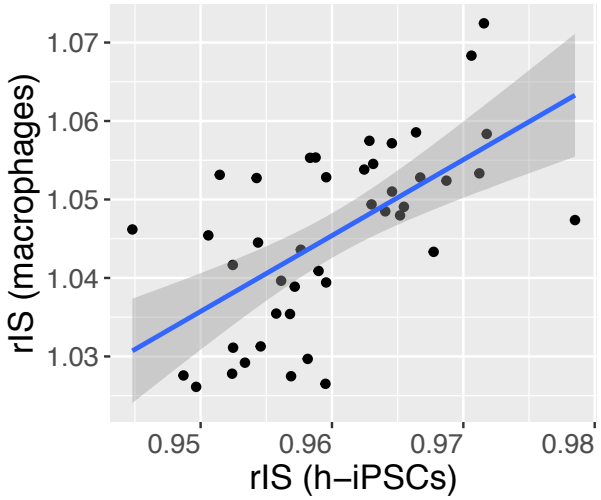


Figure 5.2: The association between rIS of iPSC-derived macrophages and their progenitor h-iPSCs (Pearson correlation $= 0.61$, p-value $= 1.3 \times 10^{-5}$).

5.3.2 Potential association between immune response and XCI level

The X chromosome is known to contain the largest group of immune related genes of the human genome (Bianchi et al. 2012b, Libert et al. 2010). In addition, the X chromosome abnormality, including both alteration of X-locus and the XCI variation, has been found associated with multiple diseases, for instance the Chronic granulomatous disease and autoimmune thyroid disease (R. Brown et al. 1993, Baehner et al. 1986, Santiwatana et al. 2018, Brix et al. 2005).

In human immune system, macrophages are widely distributed and play an indispensable role in the innate and acquired immune response (Siamon Gordon 2003, Martinez et al. 2008). For disease modeling and immunotherapy, h-iPSC derived macrophages have been seen as a promising tool and were used in many studies (H. Zhang et al. 2015, Ackermann et al. 2018, C. Z. Lee et al. 2018, Buchrieser et al. 2017).

Alasoo et al. 2018 generated iPSC-derived macrophages from HipSci lines (Kilpinen et al. 2017) and measured the expression level of iPSC-derived macrophages under different stimulus conditions. In this section, I use the data set from this

study and HipSci (Kilpinen et al. 2017) to investigate the association between XCI level and immune response.

In Alasoo et al. 2018, 86 h-iPSCs were used to derive macrophages and the methylation data is available for 43 out of these 86 h-iPSC lines in HipSci (Kilpinen et al. 2017). These 43 h-iPSCs were originated from independent female donors, 20 of which were cultured in medium Feeder Dependent (FD) while remaining 23 were cultured in medium Feeder Free (FF). Since section 5.3.1 showed the consistent of XCI level between h-iPSCs and iPSC-derived cells, here, the mIS of these 43 h-iPSCs is used as the XCI metrics for the association analysis between XCI level and immune alteration.

Alasoo et al. 2018 cultured iPSC-derived macrophages under three different stimuli conditions and one naive condition (labeled as condition A). The three stimuli conditions were: Interferon-gamma (INFg, labeled as condition B), Salmonella typhimurium (SL1344, labeled as condition C) and a combination of INFg and SL1344 (labeled as condition D). The expression level of the iPSC-derived macr-ophages under each immune condition was measured by RNA-sequencing (Alasoo et al. 2018). Raw RNA-seq counts were normalised with function *vst* in R package DESeq2 (Love et al. 2014) and only genes with mean expression in at least one of the conditions greater than 0.5 transcripts per million were kept for the analysis, making the total number of genes after filtering to 15,797 (Alasoo et al. 2018). The different expression level between a stimulus condition and the naive condition (below BA, CA, DA), is used as the representation of the immune alteration.

Before the computation of immune related expression alteration, a filtering process was carried out to remove genes of which the number of read counts was smaller than ten in each condition. After the filtering, 14,070, 14,084 and 13,549 genes were remained for the association analysis for condition BA, CA and DA, respectively. For each condition, the association test between the expression alteration and XCI level is carried out using the univariate linear model 5.2:

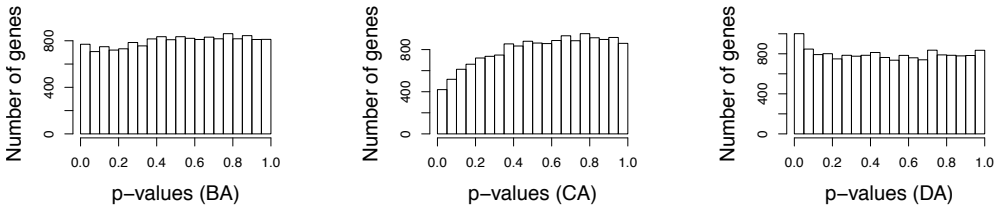
$$\text{mIS} = \log_2 \frac{gene_i \text{ (stimulus condition)}}{gene_i \text{ (reference condition)}} \tag{5.2}$$

with

$$gene_i \in list_{XA}, (X \in \{B, C, D\}).$$

Figure 5.3 shows the distribution of raw p-values of all linear models with formula 5.2 under three stimuli conditions. In this figure, an overabundance is observed for mid and high p-values for condition CA, accompanied by a valley-like curve for low p-values. This abnormal distribution of p-values follows neither normal distribution of independent tests nor the null hypothesis which would lead to abundance for low p-values.

This abnormality in the association analysis implies that some model assumptions are not met. This might be caused by confounding factors, for instance, the



(a) (b) (c)

Figure 5.3: Histogram of raw p-values for the univariate linear model for the association analysis between XCI level and immune response under three stimuli conditions. The log_2 transformed fold change of gene expression between immune stimulus condition and reference condition is used as representative of immune response and the mIS is used as representative of XCI level. The analysis uses univariate linear model (formula 5.2). a. Condition BA: alteration between stimulus condition B (INFg) and the naive condition. b. Condition CA: alteration between stimulus condition C (SL1344) and the naive condition. c. Condition DA: alteration between stimulus condition D (INFg and SL1344) and the naive condition.

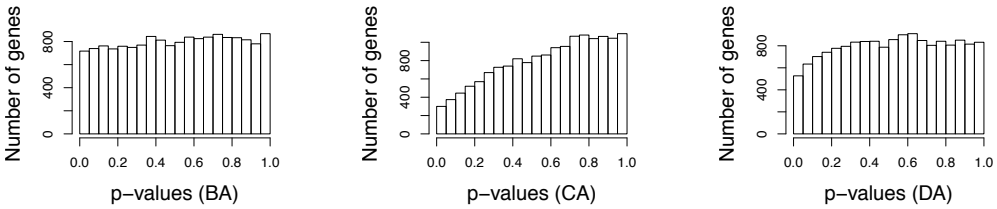
date of salmonella infection. For 43 macrophages in the study, salmonella infection was done on 27 different dates, various from 3rd June 2014 to 4th December 2015. Taking salmonella infection date as a random effect factor, linear model 5.2 is updated to mixed linear model 5.3:

$$\text{mIS} = \log_2 \frac{gene_i\,(\text{stimuli condition})}{gene_i\,(\text{reference condition})} + (1|\,\text{samonela date}). \qquad (5.3)$$

Figure 5.4 shows the distribution of p-values for model 5.3 for each immune condition. It appears that the abnormality of distribution of p-values for condition CA is not solved even when the date of salmonella infection is included in the model. Furthermore, compared to model 5.2, the distribution of p-values for condition DA is worse.

Since including the date of salmonella infection as random effect does not solve the problem of abnormal distribution of p-values, I assume that this abnormality might be caused by other unknown confounding factors.
To include all possible confounding factors in the association analysis might be unrealistic. Instead, similar as in section 2.5.2, these factors can be corrected from expression level by PEER correction method (Stegle, Parts, Piipari, et al. 2012, Stegle, Parts, Durbin, et al. 2010).

The PEER method, which is the abbreviation of probabilistic estimation of expression residuals, uses additive Bayesian network to infer hidden factors and their effects in gene expression matrix (Stegle, Parts, Durbin, et al. 2010, Stegle, Parts, Piipari, et al. 2012). Using expression data matrix as input, PEER

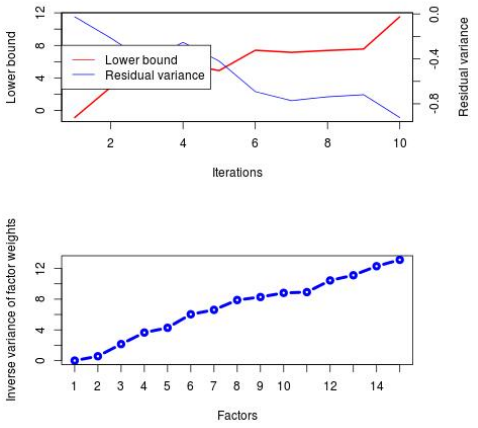


(a) (b) (c)

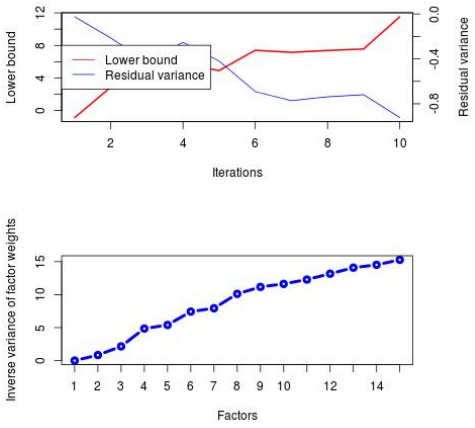
Figure 5.4: Histogram of raw p-values for mixed linear model 5.3 under three stimuli conditions. In mixed linear model 5.3, the date of salmonella infection is included as a random effect factor for the association analysis between immune-related expression alteration and the XCI level. a. Condition BA: alteration between stimulus condition B (INFg) and the naive condition. b. Condition CA: alteration between stimulus condition C (SL1344) and the naive condition. c. Condition DA: alteration between stimulus condition D (INFg and SL1344) and the naive condition.

method outputs residuals, which can be used as the corrected expression data, posterior mean, weights of the inferred confounders and precision (the inverse variance) of the weights (Stegle, Parts, Durbin, et al. 2010, Stegle, Parts, Piipari, et al. 2012). Here, I use package PEER, version 1.3 (Stegle, Parts, Piipari, et al. 2012) in R, version 3.4.0 (R Core Team 2017) to correct expression data of iPSC-derived macrophages (Alasoo et al. 2018). The PEER correction is applied on normalised RNA-sequencing counts of each condition, with number of iteration $= 1,000$ and number of factor $= 15$ (factors suggested by the tutorial of PEER package). Residual variance alteration with number of factors in the correction is shown in figure 5.5, where it can be observed that for all conditions, the residual variance is continuously decreasing after 9th iteration meanwhile factor weights smoothly increase with larger factor numbers (for conditions A and B, variance of factor weights start to smoothly increase at factor 12 and 8 respectively). These correction results reveal that PEER correction results with 15 factors and $1,000$ iterations are sufficient to remove confounding factors in these gene expression data.

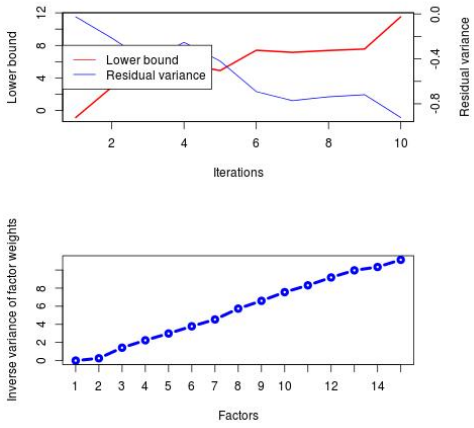
After correction, $\log_2$ transformed fold change between expression level under stimulus condition and under reference condition is computed, following by association analysis with linear model 5.2. The distribution of p-values in this linear model is shown in figure 5.6.



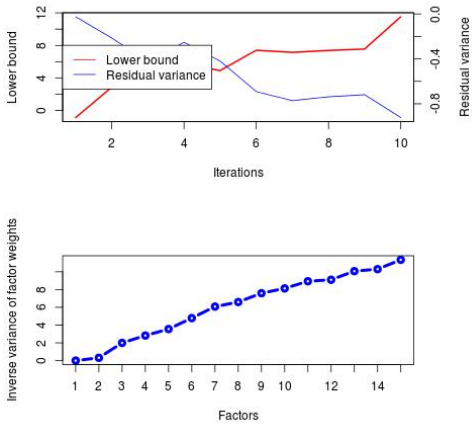
(a) Condition A



(b) Condition B

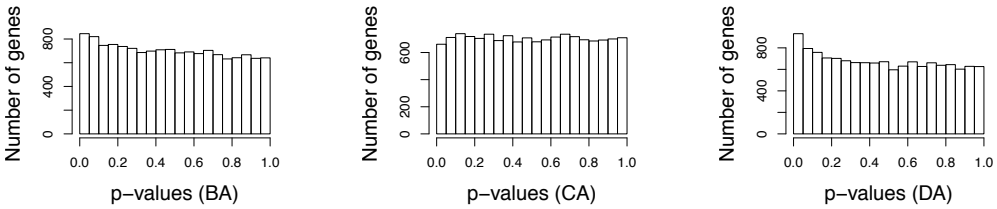


(c) Condition C



(d) Condition D

Figure 5.5: PEER correction for expression level under four immune conditions. Condition A. The naive condition (reference). Condition B. Interferon-gamma (INFg). Condition C. Salmonella typhimurium (SL1344). Condition D. The combination of INFg and SL1344.



(a) (b) (c)

Figure 5.6: Histogram of raw p-values for the univariate linear model 5.2 for the association analysis between XCI level and immune response under three stimuli conditions using PEER corrected expression data as the immune response. a. Condition BA: alteration between stimulus condition B (INFg) and the naive condition. b. Condition CA: alteration between stimulus condition C (SL1344) and the naive condition. c. Condition DA: alteration between stimulus condition D (INFg and SL1344) and the naive condition.

Accounted for multiple testing (Benjamini and Hochberg 1995), there is no genes significantly associated with XCI in any of immune conditions B, C or D (FDR $< 10\%$). Condition D (INFg + SL1344) has more genes associated with XCI level than condition A and B, as it is the condition with more immune stimuli. The association between XCI level and immune related gene alteration for six genes with smallest adjusted p-values in condition D is shown in figure 5.7.

This result reveals that although many genes have altered expression due to XCI loss (section 5.1), the magnitude of such changes is small in comparison to the changes induced by strong external stimuli such as infection. As this association study is carried out with limited sample size ($n = 43$), there is still a possibility to observe association between XCI level and immune-related activities of macrophages with bigger data set. Also, I expect to see further studies which investigate the role of XCI heterogeneity in other iPSC-derived immune cells (i.e. natural killer cells), to have a better understanding about how the X chromosome activation of h-iPSCs regulates the biological functions in their derived cells.

5.4 Discussion: broad consequences of XCI heterogeneity in h-iPSCs and iPSC-derived cells

Knowing that the XCI heterogeneity generally exists in female h-iPSCs, it is important to know what are direct consequences and to what extent are these consequences: are they limited to h-iPSCs themselves, or do they also have a power in iPSC-derived cells?

Since that XCI is the dosage compensation mechanism to balance expression level between genders (Lyon 1961, Brockdorff et al. 2015, Heard et al. 1997,

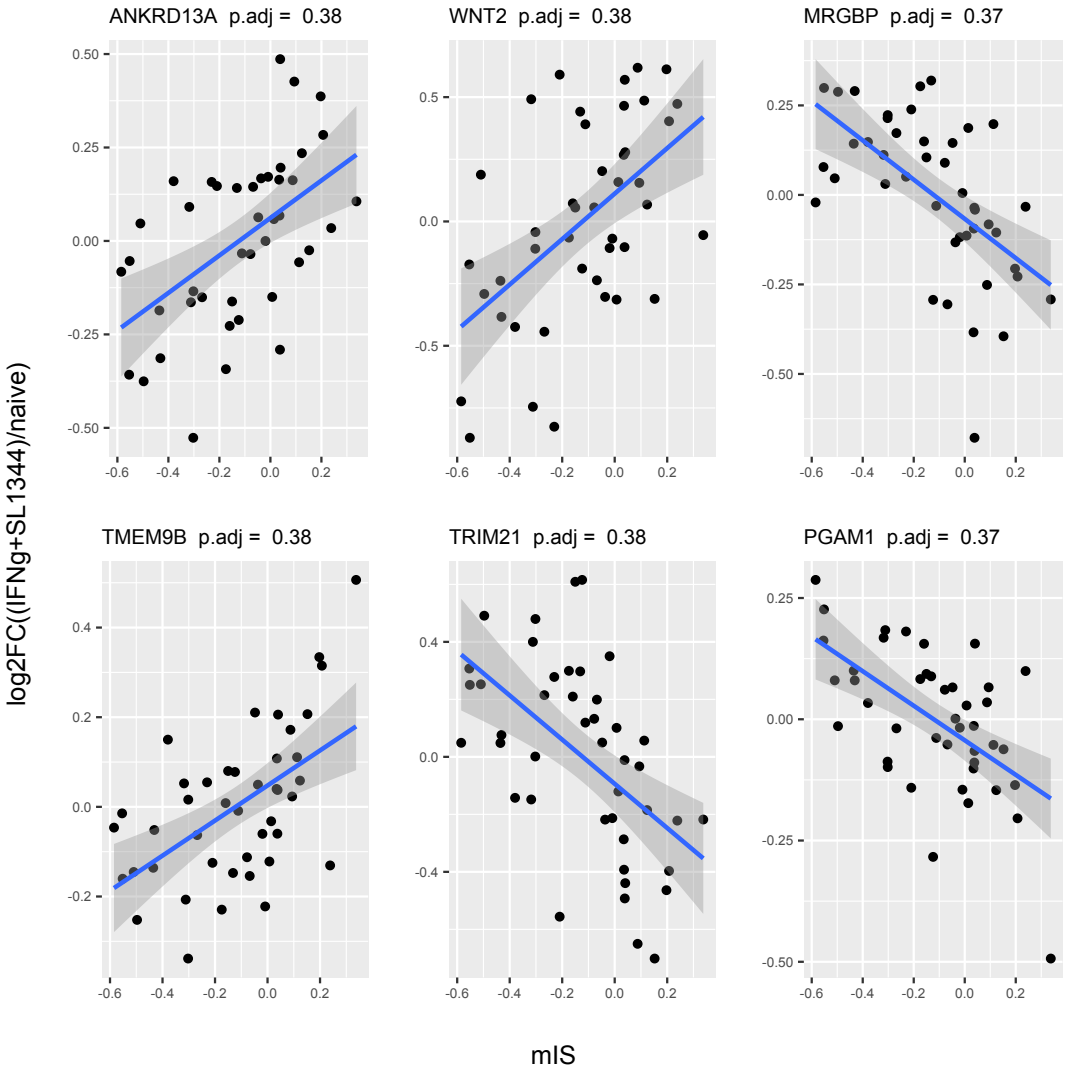


Figure 5.7: The association between XCI level (mIS) and expression alteration under condition DA for six genes with the lowest p-value in model 5.2 after multiple testing control. The x-axis is the mIS value. The y-axis is the $\log_2$ transformed fold change of PEER-corrected expression between condition D (INFg and SL1344) and condition A (reference).

Avner et al. 2001, Galupa et al. 2018), the investigation starts from the expression level of h-iPSCs, where I find a genome-wide expression level along with the XCI variation. Specifically, the consequence of XCI loss is different on the X chromosome and on autosomes: among genes which are associated with XCI, most of X-linked genes are up-regulated (97%), while the fraction of up-regulated genes on autosomes is around half (47%).
This result is consistent with previous studies which reported overexpression of X chromosome linked with XCI erosion (Bar et al. 2019, Brenes et al. 2020), proving that the loss of XCI results in an increase of X-linked expression level. Meanwhile, the consequence is not limited to the X chromosome, as shown in section 5.1, 85% of XCI-associated genes locate on autosomes, revealing that the effect of XCI loss is genome-wide.

Another interesting result is that the XCI level in h-iPSCs is inheritable by iPSC-derived cells. This result can be interpreted from two aspects. Firstly, the XCI variability is inherited by iPSC-derived cells. Therefore, in the research of disease modeling, specifically X-linked disease (i.e. Rett syndrome), the XCI level needs to be taken into account when h-iPSCs and iPSC-derived cells are used. Secondly, as one of epigenetic signatures, the inheritance of XCI shows the homogeneity between h-iPSCs and iPSC-derived cells. Considering that An et al. 2012 also presented the inheritance of gene correction from h-iPSCs to iPSC-derived neurons, there is a high possibility that iPSC-derived cells can inherit genetic features from h-iPSCs, thus it is potential to execute genetic modifications in h-iPSCs and apply iPSC-derived cells, which contain these modifications, for a certain cell therapy.

Using iPSC-derived macrophages (Alasoo et al. 2018) to explore the association between XCI level and immune response is another investigation of this thesis for the usage of h-iPSCs and iPSC-derived cells in clinical research. Knowing that X chromosome contains the largest group of X chromosome genes in human (Bianchi et al. 2012b, Libert et al. 2010), the concern is about potential immune alteration or even immune disorder with XCI variation. With the data included in this thesis, I do not observe a significant association between XCI level and the immune response. Since that the data size is till limited ($n = 43$), I am looking forward to further studies which investigate the immune related activities regarding the XCI heterogeneity in larger size of h-iPSCs and/or iPSC-derived cells.

With the work in this chapter, I demonstrate the broad consequences of XCI heterogeneity, showing as the expression alteration in h-iPSCs and as an inheritable signature to iPSC-derived cells. I hope this result will encourage further studies to explore downstream consequences of XCI heterogeneity and to explore whether these consequences have an effect in the medical usage of h-iPSCs and/or iPSC-derived cells.

Chapter 6

Validation of XCI-related analysis in h-iPSCs from LCL data set

- Can we observe XCI heterogeneity in h-iPSCs from different origins?
Yes.

Previous chapters of this thesis presented the general XCI heterogeneity, the methylation based and expression based metrics to represent the XCI level, sources of XCI variation, as well as consequences in h-iPSCs and iPSC-derived cells (Kilpinen et al. 2017). A question following these interesting discoveries is whether they are universal conclusions, regardless of origins of h-iPSCs.

HipSci (Kilpinen et al. 2017) recruited donors who are mostly UK citizens. The principle component analysis (PCA) in chapter 4 also showed the homogeneity of the donor population structure: the vast majority of h-iPSCs were clustered in the same group (163 out of 166 h-iPSCs). Meanwhile, in HipSci, all h-iPSC lines were generated from fibroblasts which were obtained from the skin biopsy of donors (Kilpinen et al. 2017).

To answer the above question, I make use of the data set generated by Banovich et al. 2018, where h-iPSC lines were generated from lymphoblastoid cell lines (LCLs) of African population Yoruba included by the 1000 Genome Project (1000 Genomes Project Consortium et al. 2015, population YRI). The major difference between this new data set and HipSci are the population of donors and the progenitor cells of h-iPSCs. Therefore, it is able to investigate whether XCI heterogeneity exists in h-iPSCs from a different population and different starting cells. Hereinafter, I use YRI-LCLs to refer to LCLs from the population YRI in the 1000 Genome Project and use LCL-iPSCs to refer to h-iPSCs which were generated from these LCLs.

The establishment of LCL samples has been well developed by Heidemarie Neitzel in 1986, with the application of Epstein-Barr virus (EBV) for the transformation from peripheral B lymphocytes to permanent growing LCLs (Neitzel

1986). Other antigens have also be used in the establishment process, for instance the Cyclosporine A (Anderson et al. 1984) and the 2-mercaptoethanol (2-ME, Steel 1972), whereas Neitze's method has been most widely applied for the establishment of LCLs. Since the initial development, LCLs have been seen with a great practical value in clinical research and for human genetics research because of relatively easy establishment, genetic stability in long passage and permanent source of re-sampling (Neitzel 1986, Wheeler et al. 2012, Talebizadeh et al. 2008, Quinn et al. 2013, Choy et al. 2008, Monks et al. 2004, Choy et al. 2008, Niu et al. 2010).

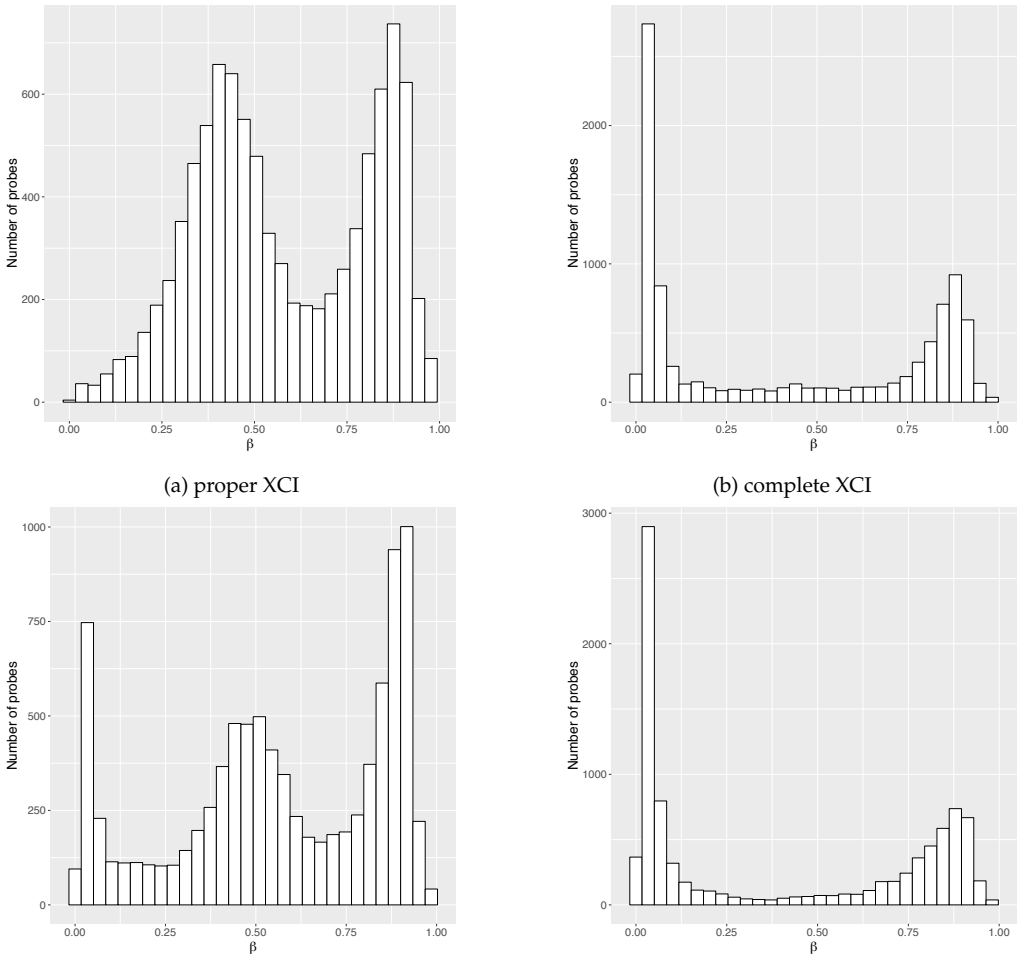
The 1000 Genome Project collected LCL samples from populations across different continents (Europe, Africa, Asia, America) and executed multiple screening on these samples (i.e. exome-sequencing and RNA-sequencing) (1000 Genomes Project Consortium et al. 2015, Sudmant et al. 2015). In the 1000 Genome Project, the subset of population YRI contains LCL samples from 32 independent females and 26 independent males.

Banovich et al. 2018 generated 58 h-iPSCs from YRI-LCL samples, of which the methylation level and RNA expression level of both autosomes and the X chromosome were measured, using the 450k methylation array and Illumina HiSeq 2500 respectively. With Banovich et al. 2018 and the 1000 Genome Project, I am able to reproduce following XCI-related analysis with LCL-iPSCs: the overview of XCI level, the consistency between XCI metrics mIS and aIS, as well as the consistency of XCI level between LCL-iPSCs and their progenitor cells (YRI-LCLs). To keep the consistency with previous analysis, the analysis in this chapter follows the same data preparation pipeline, filtering and computation process as in previous analysis.

6.1 The XCI heterogeneity in LCL-iPSCs

Similar as in chapter 2, the β value is used to represent the methylation level of the X chromosome, with the definition in formula 2.1.As described in section 2.1, the β value falls into interval $[0, 1)$. For probes on the X chromosome, when $\beta = 0$, this locus is unmethylated in all molecules; when $\beta = 0.5$, hall of molecules is methylated at this locus.

Similar as h-iPSCs in HipSci (section 2.1, figure 2.2), LCL-iPSCs also display three patterns of XCI level: the proper XCI where there is no peak at $\beta = 0.5$, the complete XCI loss where there is no peak at $\beta = 0$ and the incomplete XCI loss where there are peaks at both $\beta = 0$ and $\beta = 0.5$ (figure 6.1 a, b and c). The distribution of β in 26 male LCL-iPSCs is uniform (figure 6.1 d), which is also similar as the previous discovery with male h-iPSC lines in HipSCi (figure 2.2 d).



(a) proper XCI

(b) complete XCI

(c) Incomplete XCI loss

(d) Male LCL-iPSC

Figure 6.1: The distribution of β on the X chromosome in female and male LCL-iPSCs. a, b and c. The three patterns of distributions of β in female LCL-iPSCs, representing the proper XCI, the complete XCI loss and the incomplete XCI loss (NA18511, NA18508 and NA18520, respectively). d. An example of the uniform distribution of β in male LCL-iPSC (NA18486).

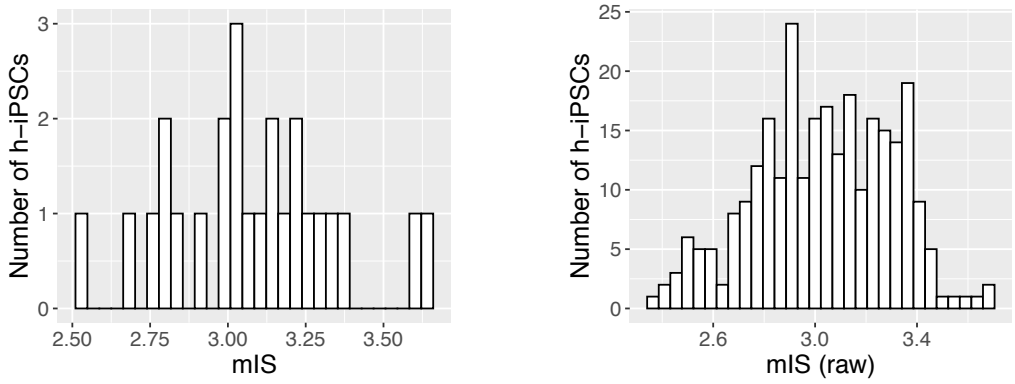
6.2 The strong correlation between mIS and aIS in LCL-iPSCs

The mIS and aIS are computed for LCL-iPSCs to have an overview of XCI level in this data set and to investigate the capacity of the expression level to estimate the XCI status in h-iPSCs. The definition and computation process of these two XCI metrics are presented in section 2.2 and section 2.3.

The methylation and RNA-sequencing data of 58 LCL-iPSCs in Banovich et al. 2018 are available on Gene Expression Omnibus (GEO), with series number GSE89895. The raw methylation data was downloaded from GEO with sub-series number GSE110544 and raw RNA-sequencing data was downloaded from SRA with series number SRP126289. The raw RNA-sequencing data of YRI-LCL samples was downloaded from the 1000 Genome Project, release version of May 2013 (1000 Genomes Project Consortium et al. 2015). All downloaded data used human reference GRCh37/hg19 (Harrow et al. 2012) for the alignment and mapping, which is the same with the genome reference in HipSci (Kilpinen et al. 2017).

6.2.1 Computation of mIS for LCL-iPSCs

In chapter 2, 9,257 probes on the X chromosome were used for the computation of mIS in h-iPSC lines. By checking the downloaded methylation data, I found that all these 9,257 probes were included in the methylation array for LCL-iPSCs so that the exact same probes as the previous analysis (section 2.2) were used for the computation of mIS for LCL-iPSC lines. The distribution of raw mIS value for 32 LCL-iPSCs generated from female YRI-LCLs is shown in figure 6.2 (a).



(a) (b)

Figure 6.2: The same range of mIS in 32 female LCL-iPSCs (a) and in 273 female h-iPSCs from HipSci (b).

The raw mIS of 32 female LCL-iPSCs falls into the interval of [2.54, 3.67], which is

similar as the interval of raw mIS in 273 female h-iPSCs from HipSci ([2.38, 3.70], figure 6.2 b). This similarity shows a consistent XCI level in female h-iPSCs regardless of origins.

In contrary with the correction process for raw mIS values in HipSci (section 2.2 and section 3.1), here, the meta data of 58 YRI-LCLs and the information in Banovich et al. 2018 show that it is unnecessary to adjust raw mIS values for LCL-iPSCs.

According to the information on the 1000 Genome Project (1000 Genomes Project Consortium et al. 2015) and Banovich et al. 2018, LCLs are from 58 independent donors and each LCL-iPSC line was generated from one single YRI-LCL. Furthermore, these 58 LCL-iPSCs were all cultured in feeder-free (FF) condition. These information reveal that the two most important confounding factors of mIS in previous analysis, namely the cell culture media and type of methylation array, do not confound the methylation data of LCL-iPSCs. Three other technical factors are recorded for the methylation array of LCL-iPSCs: the accession, the place where samples were placed (sentrix ID) and the exact place of sample on the plate (sentrix position). All 32 female LCL-iPSCs were measured with different accession so that this is not a confounding factor. In total eight sentrix IDs and eight sentrix positions were used in the methylation array. The Shapiro-Wilk normality test (J. Royston 1982, J Patrick Royston 1982, P. Royston 1995) gives a non-significant result for the raw mIS values of 32 LCL-iPSCs (p-value = 0.63), showing that these two technical factors did not have a confounding effect on the distribution of mIS in this data set. Therefore, the raw mIS is used for further analysis of LCL-iPSCs.

6.2.2 Computation of aIS for LCL-iPSCs

The aIS is defined as the average of ratio of bi-allelic expression on heterozygous positions on the X chromosome, with the computation workflow shown in figure 2.6. The RNA-sequencing data is available for 25 out of 32 female LCL-iPSCs and the data download process is described at the beginning of this chapter.

Data processing

The processing pipeline contained four steps, leading from raw fastq files to VCF files with high quality bases. The pipeline was written in bash scripts and Rmarkdown files (R, version 3.4.0, R Core Team 2017) and was executed on HPC clusters of Wellcome Trust Sanger Institute.

The alignment of fastq files

The software bwa, version 1,9.0 was used for the mapping of fastq files. The human reference genome GRCh37/hg19 (Harrow et al. 2012) was used as the reference for the mapping. The mapped bam files were then sorted and indexed

with software samtools, version 1.9.0 (Heng Li, B. Handsaker, et al. 2009).

Generation of bed file

The bed file of each sample, which contained heterozygous positions on the X chromosome, was generated from the whole genome sequencing data of YRI-LCLs, downloaded from the 1000 Genome Project website, using function *tabix* of samtools (version 1.9.0).

The variant call

The function *mpileup* of bcftools, version 1.9 (Danecek, Schiffels, et al. 2014), was called to execute the variant calling process. During this process, flag AD and DP were added, which stand respectively for the allelic depth and read depth, which stands for number of high-quality bases. These flags were used in the following analysis for the computation of aIS and for the filtering of loci.

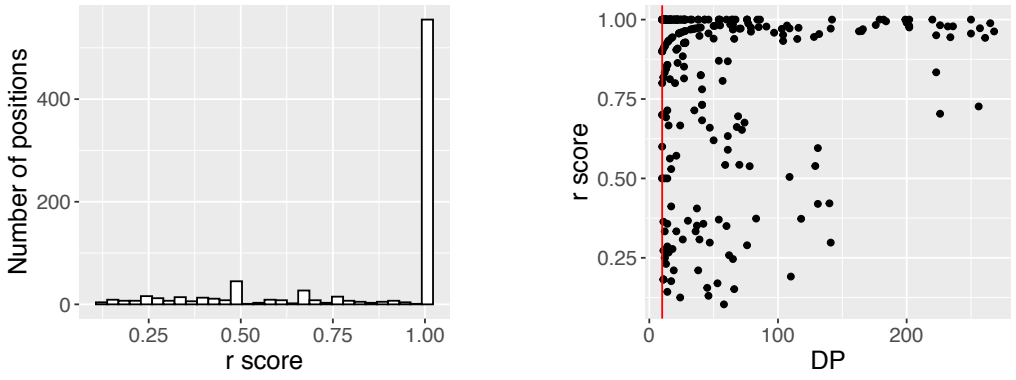
Filtering of VCF files based on the number of high quality bases (DP)

During the data processing, the bed file containing all heterozygous positions was generated for each sample. I firstly computed the r value of each position, which stands for the ratio of the number of alternative allele over the total number of alleles detected at a certain locus according to the definition in the flowchart for the computation of aIS (figure 2.6). The distribution of r values for heterozygous positions on the X chromosome show a clear enrichment of positions at r = 1.0 (figure 6.3 a). When plotting the association between the r value and the number of high quality bases (DP, flag added in VCF file), it is observed that many enriched positions with r = 1.0 had a DP value smaller than 10 (red vertical line in figure 6.3 b).

To check whether this enrichment is caused by the small number of high quality bases, the same data processing was ran on the chromosome 1 for all LCL-iPSCs. The distribution of r values on the chromosome 1, as well as the association between r values and the number of high quality bases on all heterozygous positions are shown in figure 6.4.

Similar as on the X chromosome, the enrichment of r values can be observed at r = 1.0 for heterozygous positions on chromosome 1. Since the chromosome 1 is the longest chromosome in human and is bi-allelically expressed, the distribution of r values should be close to normal distribution. Therefore, figure 6.4 show the abnormality of the enrichment of positions at $r = 0.1$. Compared with figure 6.4, figure 6.5 shows that, for sample NA18489, the removal of positions which $DP < 10$ results in a reduction of the enriched positions at r = 1.0, meanwhile, the distribution of r values appears to be closer to normal distribution.

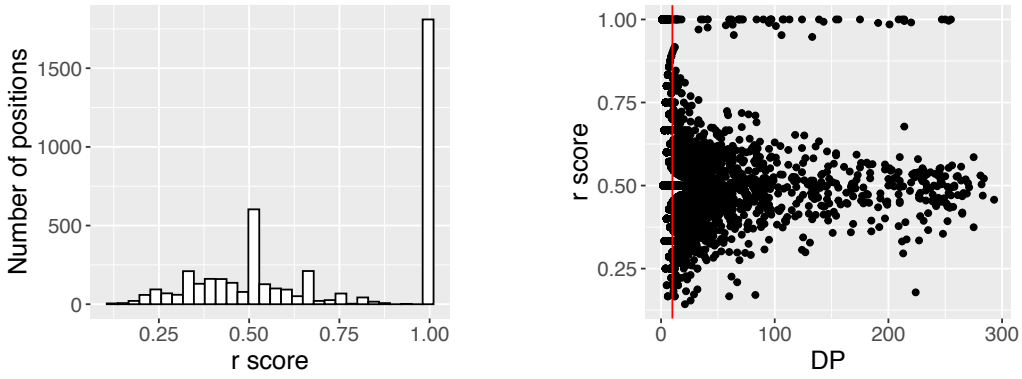
The same filtering process was conducted on the chromosome 1 of 25 female



(a) NA18489, chromosome X

(b) High quality bases (DP) and the r score

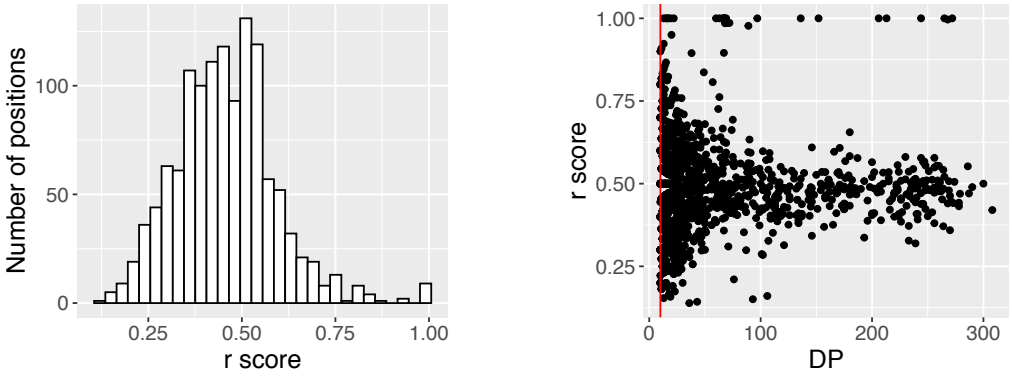
Figure 6.3: Enrichment of r = 1.0 on heterozygous positions on the X chromosome in LCL-iPSC sample NA18489. a. The histogram of r values on all heterozygous positions. b. The association between the r value and the coverage of RNA-seq on all heterozygous positions, using the number of high quality bases as the coverage (DP, an added flag in VCF files). The red vertical line refers to DP = 10.



(a) NA18489, chromosome 1

(b) High quality bases (DP) and the r score

Figure 6.4: Enrichment of r = 1.0 on heterozygous positions on the chromosome 1 in LCL-iPSC sample NA18489. a. The histogram of r values on all heterozygous positions. b. The association between the r value and the coverage of RNA-seq on all heterozygous positions, using the number of high quality bases as the coverage (DP, an added flag in VCF files). The red vertical line refers to DP = 10.



(a) NA18489, chromosome 1

(b) High quality bases (DP) and the r score

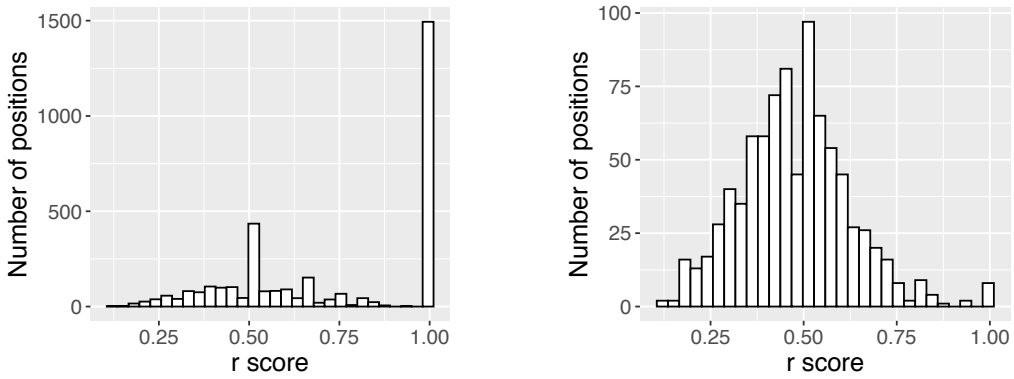
Figure 6.5: Enrichment of r = 1.0 on heterozygous positions on chromosome 1 in LCL-iPSC sample NA18489 after filtering positions of which DP is smaller than 10. a. The histogram of r values on all heterozygous positions. b. The association between the r value and the coverage (DP) of RNA-seq on filtered heterozygous positions.

LCL-iPSCs. The number of heterozygous loci on chromosome 1 was reduced from 5,350 to 1,467 on average. Meanwhile, the enriched values were significantly reduced in the histogram of r values, similar as figure 6.5 (a). I specifically checked sample NA18912, which contain the largest number of heterozygous loci on the chromosome 1, and found that most of heterozygous loci were mapped with low number of high quality bases, resulting in an obvious enrichment of r at 1.0, whereas this enrichment was removed by the filtering (figure 6.6).

The same filtering process was applied on the chromosome X for all female LCL-iPSCs. The filtering process reduces the number of X-related heterozygous positions from averagely 1,076 to 245. The minimum and maximum number of X-related heterozygous loci are 425 and 2,083 before filtering, whereas 50 and 485 afterwards. Using LCL-iPSC of NA18489 as an example, the comparison of X-related r values before and after the filtering process is shown in figure 6.7.

The computation of aIS

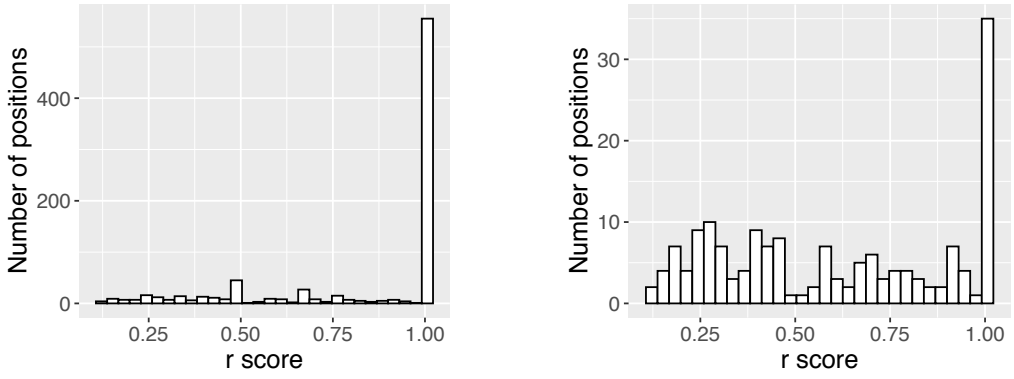
The aIS is computed with X-related r values after the filtering process, following the process shown in figure 2.6. Similar as section 2.3.1, two patterns were used: the average and the median of filtered heterozygous positions on the X chromosome. The aIS values computed by these two patterns show a good association (figure 6.8 b), which was also observed in h-iPSC lines in HipSci (figure 6.8 d). As discussed in section 2.3.1, using the median for aIS-computation would ignore the small number of loci with bi-allelic expression when the vast majority is mono-allelically expressed. To take the activation status of the entire X chro-



(a) NA18912 (chr1, before filtering)

(b) NA18912 (chr1, after filtering)

Figure 6.6: The comparison of the distribution of r values on chromosome 1 for LCL-iPSC sample NA18912 before (a) and after (b) the removal of heterozygous positions of which DP is smaller than 10.



(a) NA18489 (X, before filtering)

(b) NA18489 (X, after filtering)

Figure 6.7: The comparison of the distribution of r values on the X chromosome for LCL-iPSC NA18489 before (a) and after (b) the removal of positions of which DP is smaller than 10.

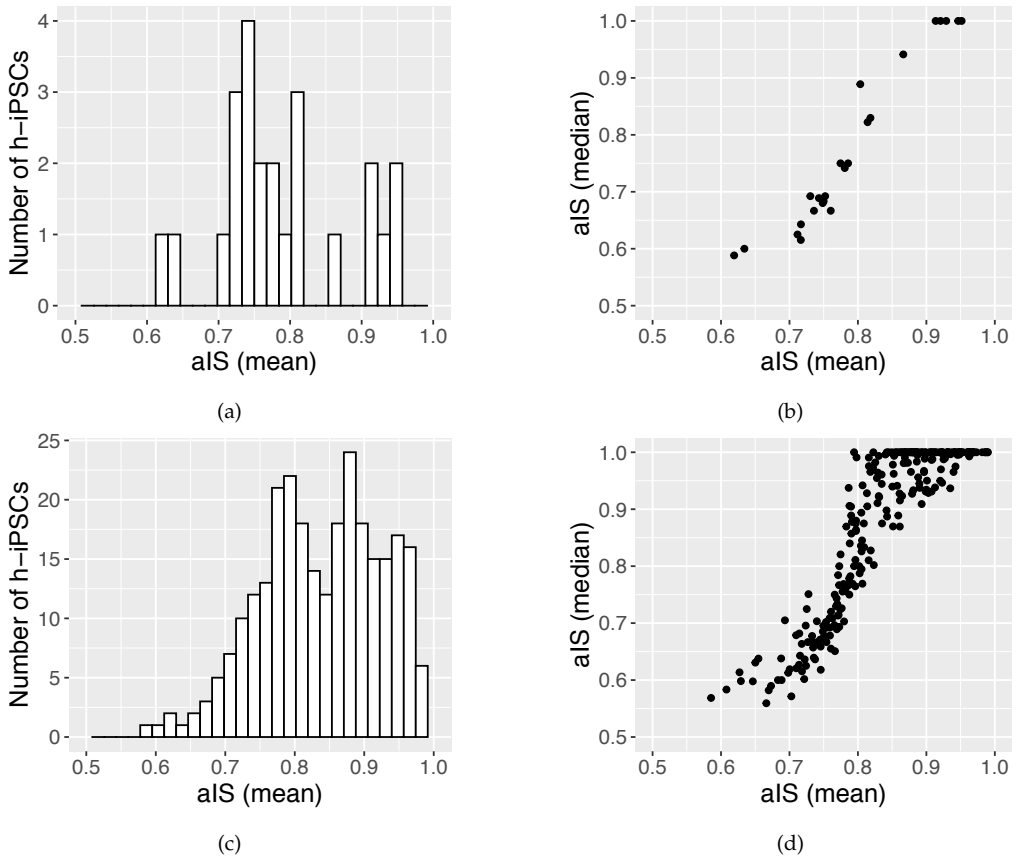


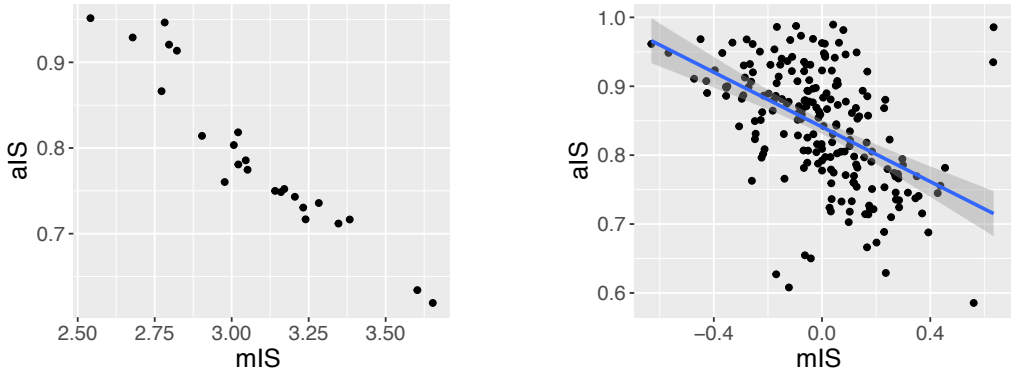
Figure 6.8: Summary of aIS in HipSci and in the LCL-iPSC data set. a. The distribution of aIS in 25 female LCL-iPSCs. b. The association between aIS using two computation patterns in 25 female LCL-iPSCs. c. The distribution of aIS in 273 female h-iPSCs in HipSci. d. The association between aIS using two computation patterns in 273 female h-iPSCs in HipSci.

mosome into account, the average-computed aIS is used as the XCI metrics. In the following text, the aIS refers to the average-computed aIS.

The distribution of aIS for 25 female LCL-iPSCs is shown in figure 6.8 (a). As a comparison, the distribution of aIS in 273 female h-iPSCs from HipSci is shown in figure 6.8 (c). The majority of h-iPSCs from HipSci has higher aIS values than LCL-iPSCs: the mean and median for h-iPSCs are 0.84 and 0.85, respectively; while for LCL-iPSCs are 0.79 and 0.76, respectively. In general, h-iPSCs from HipSci are closer to mono-allelic expression than LCL-iPSCs.

6.2.3 The strong association between mIS and aIS

For 25 female LCL-iPSCs of which both methylation and RNA-sequencing data are available, the Pearson correlation between mIS and aIS reaches -0.96, much higher than the correlation in 205 female h-iPSCs which were randomly selected by one line per donor in HipSci (Pearson correlation = -0.5), shown in figure 6.9.



(a) (b)

Figure 6.9: The correlation between XCI proxies mIS and aIS. a. In 25 LCL-iPSCs (Pearson correlation = -0.96). b. In 205 h-iPSCs which were randomly selected by one line per donor in HipSci (Pearson correlation = -0.5). The sub-figure b was shown in chapter 2, in figure 2.9 (a).

Different factors might lead to this observation of high correlation between XCI metrics, for instance, the limited sample size, cell culture time and the source of LCL-iPSCs.

Only 25 LCL-iPSCs are available with both methylation and expression level data, thus this high correlation would not be seen as a population-based conclusion. According to Banovich et al. 2018 (supplementary information), LCL-iPSCs were cultured at least three passages before being collected for analysis, whereas each passage took averagely seven days (Guideline of Handing human iPSCs by the Cedars-Sinai hospital, USA), making cell culture time starting from

approximately 21 days. The detailed passage information is available for 11 LCL-iPSCs, where the maximum passage number was 14, the minimum number was 5 (mean = 6), corresponding to cell culture time from 35 days to 98 days (mean = 42 days). Compared to lines in HipSci where cell culture time varied from 24 days to 240 days (mean = 77 days; median = 72 days), cell culture time is shorter for most LCL-iPSCs. As presented in chapter 3, the expression of XIST drops sharply at day 50 while XCI level does not follow (figure 3.8). With a lack of culture time for all 25 LCL-iPSCs, it is not able to confirm the time-effect on the high consistency between two XCI metrics for LCL-iPSCs.

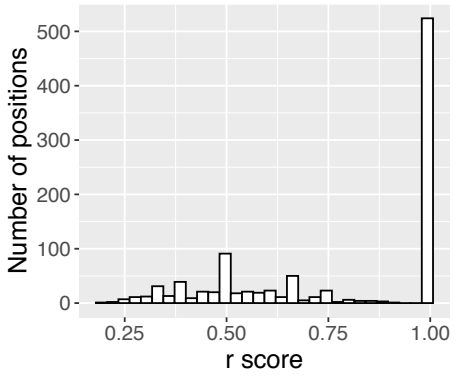
At last, different cells were used as the origins for the generation of human iPSC lines in these two data sets: fibroblasts in HipSci (Kilpinen et al. 2017) and EBV-transformed LCLs in Banovich et al. 2018. Even though fibroblasts have been widely used as starting cell of h-iPSCs in previous studies, the genetic similarity between fibroblasts and derived h-iPSCs is still unclear (Pomp et al. 2011, Tchieu et al. 2010, Anguera et al. 2012). Furthermore, the study by Rajesh et al. 2011 used two EBV transformed LCLs and derived two clonal iPSCs from each LCL sample, where the EBV antigen was detected in LCLs but not LCL-iPSCs, revealing that LCL-iPSCs contain genetic variations from their progenitor cells.

6.3 The random pattern of XCI alteration in the generation of LCL-iPSCs from LCLs

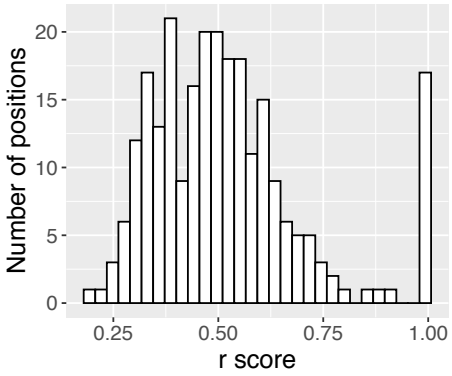
Chapter 5 presented that XCI level can be stably inherited by iPSC-derived cells (figure 5.2). Here, I investigate whether h-iPSCs could inherit XCI level from the starting cells: the correlation between XCI level of LCL-iPSCs and of YRI-LCLs. In this association analysis, aIS is used as XCI metrics.

The RNA-sequencing data of YRI-LCLs, including both autosomes and chromosome X, were downloaded from the 1000 Genome Project (1000 Genomes Project Consortium et al. 2015), version 201305 and were mapped to human reference hg19/GRCh37 (Harrow et al. 2012). The same pipeline was used for the alignment and the variant calling process for YRI-LCLs, where the enrichment of heterozygous positions can be observed again at r = 1.0 in the chromosome 1 and the chromosome X in female LCLs before filtering (figure 6.10, a and c). The filtering process with removing heterozygous positions of which DP < 10 helps the distribution of r value closer to normal distribution in chromosome 1 (figure 6.10 d) and removes most of the enriched r values at 1.0 for the X chromosome (figure 6.10 b).

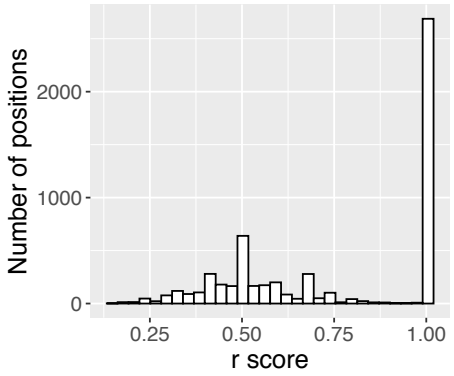
For 25 female YRI-LCLs, the distribution of average-computed aIS is shown in figure 6.11 (a) and the association between aIS computed by two patterns is shown in figure 6.11 (b). For YRI-LCLs with relatively high average-computed aIS (aIS $>$ 0.85), the median-computed aIS do not concentrate at value 1.0, but close to the average-computed value (figure 6.11 b).



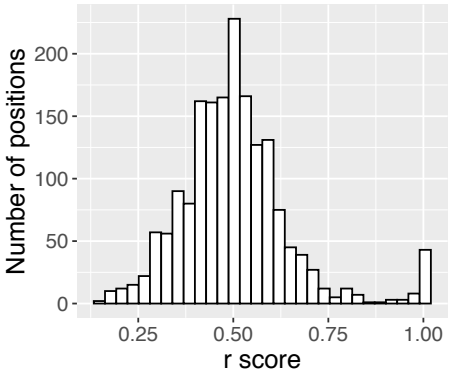
(a) NA18489 (YRI-LCLs, chrX, before filtering)



(b) NA18489 (YRI-LCLs, chrX, after filtering)



(c) NA18489 (YRI-LCLs, chr1, before filtering)

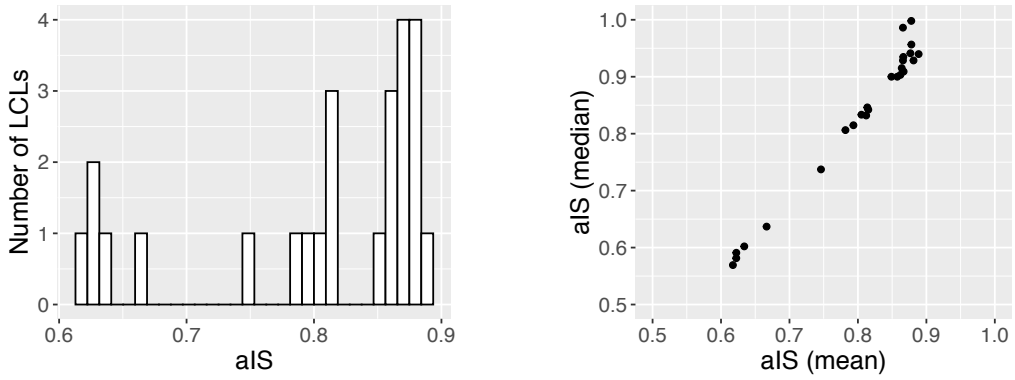


(d) NA18489 (YRI-LCLs, chr1, after filtering)

Figure 6.10: The Comparison of distribution of r values before (left) and after (right) filtering process with the removal heterozygous positions of which DP is smaller than 10 in chromosome X (up) and in chromosome 1 (bottom) for sample NA18489.

This observation leads to an assumption of different XCI pattern in LCLs and in LCL-iPSCs: it is less common in LCLs than in LCL-iPSCs to have a few bi-allelic expression when the majority of X chromosome is mono-allelically expressed, thus the XCI is more complete in YRI-LCLs (starting cells) than in LCL-iPSCs.

Besides the different distribution of aIS in LCLs and LCL-iPSCs, there is poor association between aIS of these two types of cells (Pearson correlation = 0.24, figure 6.12). Figure 6.12 can be interpreted in two parts. Firstly, points which locate in the top-middle of the figure show that LCL-iPSCs either inherit the XCI level from the YRI-LCLs or contain a XCI loss. This XCI variability was observed in h-iPSCs in HipSci (chapter 2) and both patterns were reported by previous studies (Mekhoubad et al. 2012, Anguera et al. 2012, Pomp et al. 2011, Marchetto et al. 2010). Secondly, points which locate on the bottom-middle of the figure display a slight higher level of XCI compared to their starting cells,



(a) (b)

Figure 6.11: Summary of aIS in 25 female YRI-LCLs. a. The distribution of aIS. b. The association between average-computed and median-computed aIS.

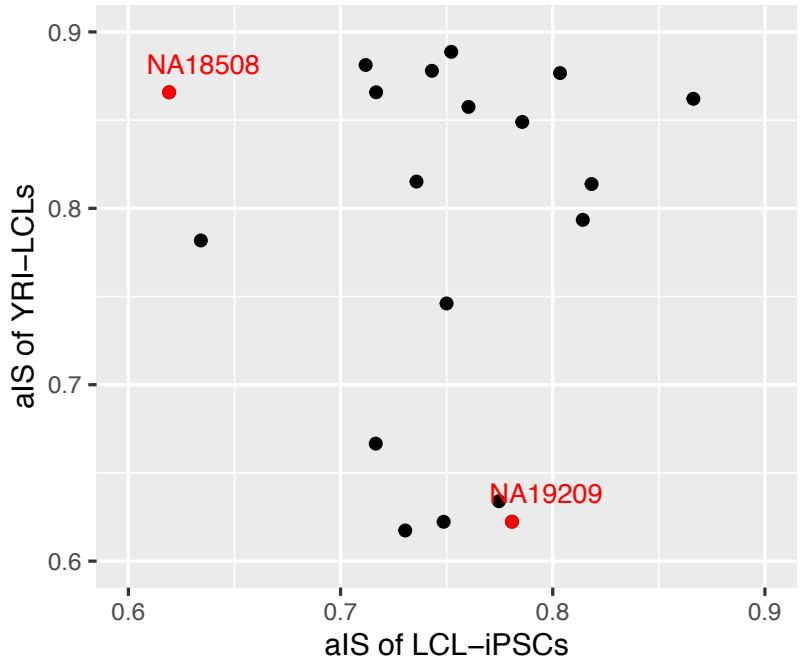


Figure 6.12: Association between aIS in LCLs and in LCL-iPSCs in all 25 female samples with two samples labeled in red for further analysis (NA18508 and NA19209).

meaning that they regain the XCI during the programming of LCL-iPSCs.

To further explore the alteration of aIS in the generation of LCL-iPSCs, I selected two samples (NA18508 and NA19209) and studied their aIS-distribution on all filtered heterozygous positions on the X chromosome in both YRI-LCLs and LCL-iPSCs. Figure 6.13 shows that sample NA18508 has vast mono-allelic expression level of the X chromosome in YRI-LCLs while almost bi-allelic expression level in LCL-iPSCs, meaning that the XCI level of this sample altered from almost-appropriate to complete loss during the generation of h-iPSC lines. However, the alteration pattern of XCI in sample NA19209 is in reverse: it displays almost complete XCI loss in YRI-LCLs (aIS close to 0.5) but regain high proportion of XCI (aIS close to 0.8) in LCL-iPSCs.

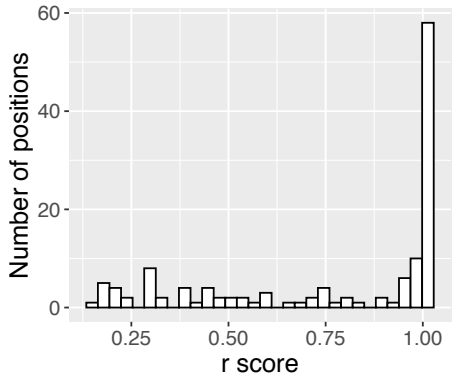
This observation points out that during the generation of h-iPSCs from LCLs, the regulation of XCI happens in a random manner: h-iPSCs have the possibility to regain proper XCI or to lose XCI at a various level. This observation also supplements the discovery in Rajesh et al. 2011 (briefly discussed in section 6.2.3) that unlike EBV transformed LCLs, the LCL-iPSCs did not contain EBV antigens, showing that the generation of h-iPSCs from LCLs was accompanied by the genetic variation at the X chromosome and at the expression level.

Lacking the RNA-sequencing data of fibroblasts in HipSci, it is impossible to investigate the regulation of XCI during the generation of h-iPSCs from fibroblasts. Further validation researches with large number of h-iPSCs derived from fibroblasts and/or from LCLs will be very meaningful for the understanding of XCI regulation during the reprogramming of h-iPSCs and I am looking forward to seeing more biological studies to investigate the potential XCI loss with large scale h-iPSC data set.

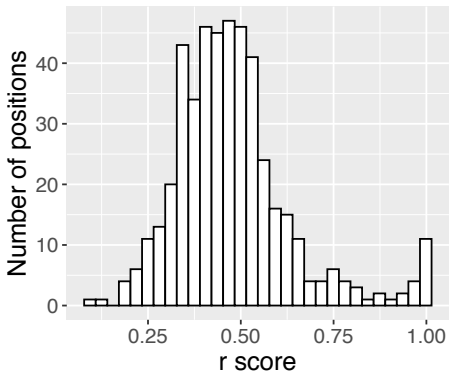
6.4 An inspiration: the XCI heterogeneity is more likely to be caused by the loss of XCI, instead of by the reactivation of the entire X chromosome

To summarize, in this chapter I demonstrate the existence of XCI heterogeneity with h-iPSCs from different progenitor cells and populations and present that methylation level and expression level are good representations of XCI for h-iPSCs. In addition, there are innovative discoveries regarding the pattern of XCI in h-iPSCs.

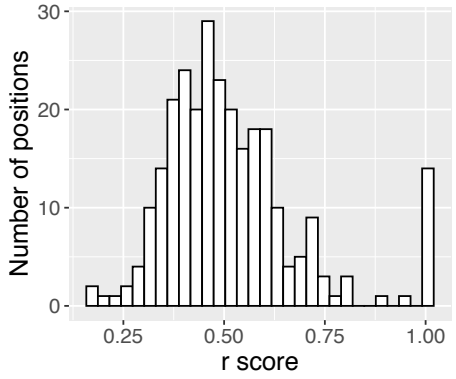
It has been well studied and clear that during the generation of iPSCs from mice (below m-iPSCs), the inactive X chromosome get reactivated thus m-iPSCs contain two active X chromosomes (Okamoto et al. 2004, Heard et al. 1997, G. Fan et al. 2011). However, about the regulation of XCI during the generation of h-iPSCs, there is not yet an conclusion: some scientists think that there is also a



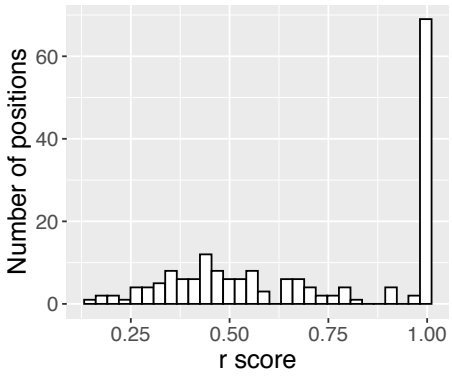
(a) NA18508 (YRI-LCL, chrX)



(b) NA18508 (LCL-iPSC, chrX)



(c) NA19209 (YRI-LCL, chrX)



(d) NA19209 (LCL-iPSCs, chrX)

Figure 6.13: The alteration of aIS in the generation of LCL-iPSCs from YRI-LCLs in sample NA18508 and sample NA19209. a and b. NA18508 has vast mono-allelic expression in YRI-LCLs (starting cells of h-iPSC generation) and has approximate bi-allelic expression in LCL-iPSCs, showing a loss of XCI level. c and d. NA19209 has bi-allelic expression in YRI-LCLs but close to mono-allelic expression in LCL-iPSCs, meaning that XCI is re-established during the generation of LCL-iPSCs.

reactivation of the X chromosome (Barakat et al. 2015, Vacca et al. 2016, Kim, Hysolli, and Park 2011, Tomoda et al. 2012), while others assume that there is a loss of XCI, which might result from the culture (Kim, Hysolli, Tanaka, et al. 2014, Mekhoubad et al. 2012, Anguera et al. 2012, Pomp et al. 2011).

With summary of results in this chapter and in chapter 2, here are my assumptions about the XCI regulation in the reprogramming of h-iPSCs: a) h-iPSCs inherit the XCI level from the donor; b) the variation of XCI in h-iPSCs results from the loss of XCI but not from the reactivation of the entire X chromosome.

The assumption a) was inspired by chapter 3 where the donor effect was identified to have an important effect in XCI variation and by chapter 4, where one variant region on chromosome 1 (section 4.2) and the rs3790598-MOV10 path were found promising as XCI determinants.

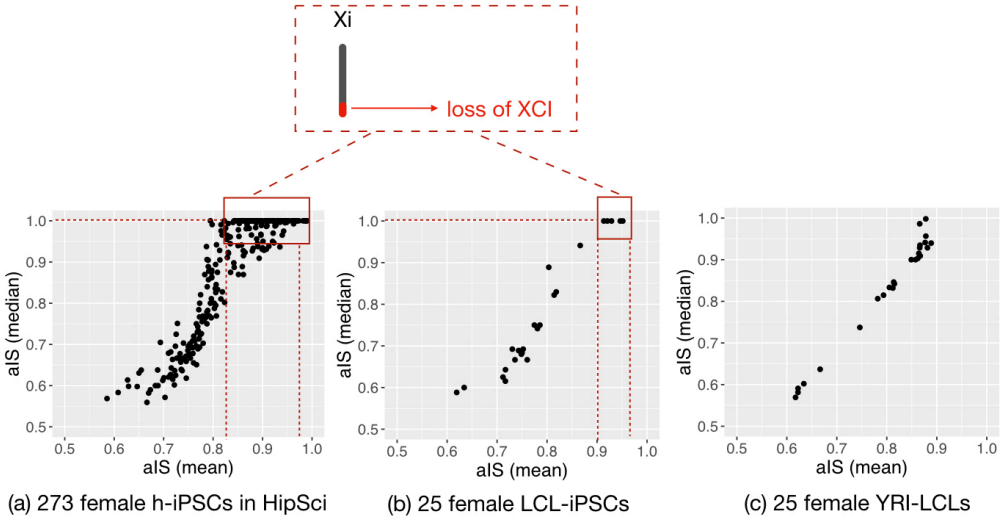


Figure 6.14: The aIS in two data sets included in this thesis reveals the pattern of XCI regulation in the programming of h-iPSCs. a and b. For 273 female h-iPSCs in HipSci and for 25 female LCL-iPSCs, a 'tail' is observed in the scatter plot of aIS computed by two patterns. c. For 25 YRI-LCLs, which are starting cells of LCL-iPSCs, there is no 'tail' in the scatter plot of computed aIS values.

Here, I explain my assumption b) with figure 6.14. For 273 female h-iPSC lines in HipSci (Kilpinen et al. 2017) and 25 female LCL-iPSCs generated by Banovich et al. 2018, a 'tail' can be observed in the scatter plot of aIS values computed by the average or the median of r values (definition in section 2.3, shown with figure 2.6). This 'tail' shows h-iPSC lines of which the median-computed aIS is 1.0 while the average-computed aIS is in the range of $[0.8, 1.0]$ (for h-iPSCs, figure 6.14 a) or of $[0.9, 1.0]$ (for LCL-iPSCs, figure 6.14 b). This difference of aIS values computed by the two patterns reveals the situation of the X chromosome: the vast majority is mono-allelic expressed while a small part is bi-allelic expressed.

In other words, h-iPSC lines which are the 'tail' display a slight loss of XCI while they are close to the proper XCI status.

Two other conclusions might be drawn with figure 6.14. Firstly, when comparing the scatter plot of aIS values in LCL-iPSCs and YRI-LCLs (figure 6.14 b and c), the 'tail' does not exist in YRI-LCLs which are starting-cells of LCL-iPSCs, meaning that the slight XCI loss is probably gained during the programming or culture of lines. Secondly, the 'tail' only exists at the mono-allelic end of the scatter plot (close to 1.0 on the x-axis of figure 6.14). In the situation that the reactivation of the X chromosome happens during the programming of h-iPSCs and follows by the inactivation process, it would be possible to observe another 'tail' at the bi-allelic end of the scatter plot (close to 0.5 on the x-axis of figure 6.14), at least in the data set of HipSci with 273 female lines.

Among the 25 LCL-iPSCs studied in this chapter, 4 lines show a capacity to regain the XCI from their progenitor cells (figure 6.4 c, middle-bottom points). Because of the limited sample size, it is hard to tell whether this is a single-line variability or whether this is a recovery of XCI in LCL-iPSCs. Due to the lack of aIS values for fibroblasts, this thesis is not able to investigate whether this potential pattern of XCI recovery also takes place in HipSci. I expect to see more studies with XCI level in both starting cells and h-iPSCs, which will be helpful to clarify the XCI regulation during the generation of h-iPSCs.

Discussion: XCI heterogeneity - the knowns and unknowns

There has been a long time argue about the XCI level in h-iPSCs: some studies assume a reactivation of the X chromosome (Barakat et al. 2015, Kim, Hysolli, Tanaka, et al. 2014), while others assume a loss of XCI during the generation of h-iPSCs (Mekhoubad et al. 2012, Anguera et al. 2012, Tchieu et al. 2010, Brenes et al. 2020, Nazor et al. 2012).

The motivation for this thesis was to investigate XCI with the large-scale female h-iPSCs from HipSci (Kilpinen et al. 2017). In previous chapters, I presented the XCI heterogeneity in 273 female h-iPSCs from 205 donors and investigated major concerns regarding this heterogeneity: how similar are sibling-lines compared with h-iPSCs from different donors; how do single cells display XCI status; which factors might be sources of this variability and what are the consequences. Specifically, I presented the donor effect in the XCI level of h-iPSCs, with which I assumed that the XCI level is mainly inherited from the donor and is slightly lost in the programming and culture of h-iPSCs.

The major advantages of this thesis compared to previous studies

The first and biggest advantage of this thesis is the large number of enrolled donors - in total 205 female donors, including 170 healthy donors and 35 patients (21 donors had Bardet-Biedl syndrome and 14 donors had neonatal diabetes). Compared with previous studies which included fewer than 15 donors or fibroblasts on average (Kim, Hysolli, Tanaka, et al. 2014, Trokovic et al. 2015, Pomp et al. 2011, Anguera et al. 2012), the population-level data set helps to improve our understanding of XCI in h-iPSCs. The second advantage is the massive number of h-iPSC lines in the analysis. With 273 female h-iPSCs and 219 male h-iPSCs, this thesis avoids the potential bias introduced by small sample size (in general smaller than 50 h-iPSC lines in Mekhoubad et al. 2012, Tchieu et al. 2010, Pomp et al. 2011, Anguera et al. 2012). Thirdly, HipSci is one of the largest data bank of h-iPSCs generated by a single institute in the world. Since different experimental settings bring noise and bias in h-iPSCs (Newman et al. 2010, Volpato et al. 2018, Rao et al. 2012), this uniform data source guarantees a well controlled input for research.

XCI heterogeneity in h-iPSCs

XCI heterogeneity is clearly shown with 273 female h-iPSCs: the vast majority of h-iPSC lines displays a varying degree of XCI level while 4 lines (1%) have complete XCI loss. This XCI heterogeneity is also observed in 32 h-iPSCs generated from LCLs (Banovich et al. 2018), originating from an African population (YRI) in the 1000 Genome Project (1000 Genomes Project Consortium et al. 2015), which is different from HipSci lines in both population of donors and types of progenitor cells. These results demonstrate that XCI heterogeneity exists in general in h-iPSCs, regardless of origin.

New discoveries regarding XCI in h-iPSCs

The expression level of h-iPSCs can be used as XCI metric, which has good association with the methylation-based XCI metric. This result can facilitate the estimation of XCI level in female h-iPSCs in laboratories without setting up a methylation array, since the expression level is commonly used in the pluripotency assay (Müller et al. 2011).

The sharp drop of XIST expression in cell culture (approximately day 50) is a very interesting observation in this thesis. This 'on/off' mode of XIST reveals a potential time-related mechanism in the expression of XIST and also shows that XIST is not a perfect marker of XCI status, especially for h-iPSCs in long cell culture.

The usage of different cell culture media results in stratification of XCI level and of pluripotency score of h-iPSC lines (Kilpinen et al. 2017). Therefore, for scientists who would like to reproduce a study or to compare their own studies with an existing result, I recommend maintaining the same cell culture medium for the generation of h-iPSCs.

The XCI variation of h-iPSCs is inherited by iPSC-derived macrophages (Alasoo et al. 2018, $n = 43$, Pearson correlation = 0.61, p-value $= 1.3 \times 10^{-5}$), revealing that when using iPSC-derived cells for disease modeling or for drug development, the XCI level should be taken into account to avoid potential bias in gene expression.

Functional consequences of XCI loss

XCI variation is associated with genome wide expression alteration in h-iPSCs, with 15% of affected genes on the X chromosome and 85% on autosomes. Similar to Bar et al. 2019 (using h-ES cells) and to Brenes et al. 2020 (using h-iPSCs), XCI loss is accompanied by an up-regulation of X-linked genes. The XCI loss has a random effect on expression of autosomes: the fraction of up- and down-regulation is almost the same with XCI loss in h-iPSCs (47% up-regulation).

Different to studies which reported XCI loss with culture time (Mekhoubad et al. 2012, Anguera et al. 2012, Nazor et al. 2012), I showed that the overall XCI level in h-iPSCs is not associated with culture time in the long term (min = 24 days, max = 240 days, mean = 78 days).

One more step towards the clinical application of h-iPSCs

The h-iPSC lines have been seen as a very promising tool for disease modelling, cell therapy and drug development (Galupa et al. 2018, Liang et al. 2013, Imaizumi et al. 2012, Y. Li et al. 2018, S. P. Paşca et al. 2011, Y.-T. Lin et al. 2018). With enormous amount of research, it is not only important to know the biological process or mechanism of XCI in h-iPSCs, but also to know consequences of XCI heterogeneity in the application of h-iPSCs in clinics.

By the end of my PhD, I ask myself this question: what are contributions of my thesis for the practical use of h-iPSCs?

Firstly, I would like to highlight that, for scientist who use h-iPSCs and/or iPSC-derived cells for disease modelling, especially X-linked diseases (i.e. Rett syndrome) and immuno-therapies (i.e. autoimmune throid disease), it is essential to include XCI level as a covariate in the research because it directly results in gene expression alteration and might have immune-related consequences. Secondly, in clinical research, it is common that multiple h-iPSCs are generated from the same donor due to a limited number of patients (Avner et al. 2001, Nazor et al. 2012, Mekhoubad et al. 2012, Anguera et al. 2012, Pomp et al. 2011). This thesis demonstrates the similarity of XCI level in h-iPSCs generated from the same donor. Nevertheless, I suggest that scientist should be very careful in this situation and check the XCI level of all h-iPSC lines before including them in the clinical research. Thirdly, for scientists and laboratories who have the expression level but not the methylation level of their h-iPSCs, a fast estimation of XCI level can be achieved by the computation of expression-based XCI metrics, namely aIS and rIS. I recommend to not use XIST expression as the marker for XCI, since this thesis shows the gap between these two factors and presents a sharp drop of XIST expression in cell culture.

What are remaining questions?

First of all, what is the XCI status at the single cell level? Even though I presented that 84 single cells of joxm_1 have largely the same XCI level, a larger single-cell data set is essential to investigate this question. Secondly, can h-iPSCs regain the XCI when progenitor cells display XCI loss? In the data set of Banovich et al. 2018, 4 out of 25 h-iPSCs showed higher X-inactivation level than their progenitor cells (LCLs). This observation and the pattern of XCI regulation in h-iPSCs need to be studied with a larger sample size. Thirdly, what is the reason that XIST drops sharply during cell culture and what leads to the variable X-methylation level when XIST is not expressing? Considering the gap

between XIST expression and XCI level, is there a different regulation pattern of XCI besides XIST in h-iPSCs? Fourthly, what are consequences of XCI variation in the application of h-iPSCs in disease modelling (i.e. Rett syndrome) and are there XCI-related side-effects due to variable methylation and/or expression level on the X chromosome?

There is never an end of scientific exploration, I hope my work and my thesis can give inspirations to other scientists in the relevant field and I am excited to follow further studies about the XCI in h-iPSCs, as well as to see more applications of h-iPSCs in biological and clinical research.

Appendix A Author's major work and publications

1. Manuscript which presents the work of this thesis:

Nan Li, Daniel Gaffney, Angela Teresa Filimon Gonçalves. The landscape of X chromosome inactivation in human induced pluripotent stem cells
(manuscript in preparation)

2. Manuscript which is related to the work of this thesis:

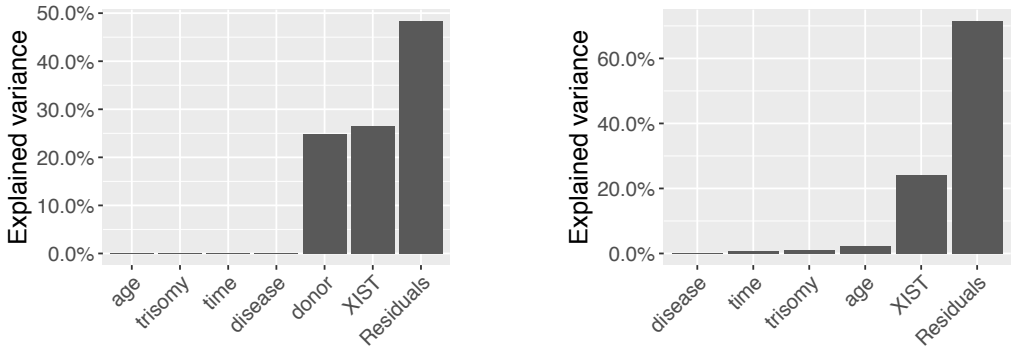
Yan Zhou, Eugene Kwa Jing, Nan Li, Angela Teresa Filimon Gonçalves, Leopold Parts. Germline variation in human gene essentiality
(ongoing collaboration and manuscript in preparation)

In addition to the work with human induced pluripotent stem cells, the author of this thesis also worked intensively on the development and application of statistical methods in drug screen projects.

3. Nan Li, Roman Schefzik, Angela Teresa Filimon Gonçalves. The application of stepwise Elastic Net regression with pairwise interaction in biomarker selection with multiple data types
(manuscript in preparation)

4. Jonatan Zorea, Manu Prasad, Limor Cohen, Nan Li, Roman Schefzik, Susmita Ghosh, Barak Rotblat, Benedikt Brors and Moshe Elkabets. IGF1R up regulation confers resistance to isoform-specific inhibitors of PI3K in PIK3CA-driven ovarian cancer. *Cell death and disease* 9.10 (2018): 1-12.

Apprendix B Supplementary Figures



(a) (b)

Figure B.1: Variance component analysis (VCA) using aIS as XCI metric, without CNA segments (included factors: cell culture time, age and health condition of donor, XIST expression level, trisomy situation of X chromosome). a. First model is fit for all 273 female h-iPSC lines, including donor as random effect factor. The first model identifies XIST expression (> 25%) and donor (25%) as most important factors for XCI heterogeneity. b. Second model is fit for 205 female h-iPSCs randomly selected from the initial data set by one line per donor. This model identifies XIST expression as the only important factor which explains slightly more than 20% of variance in XCI level.

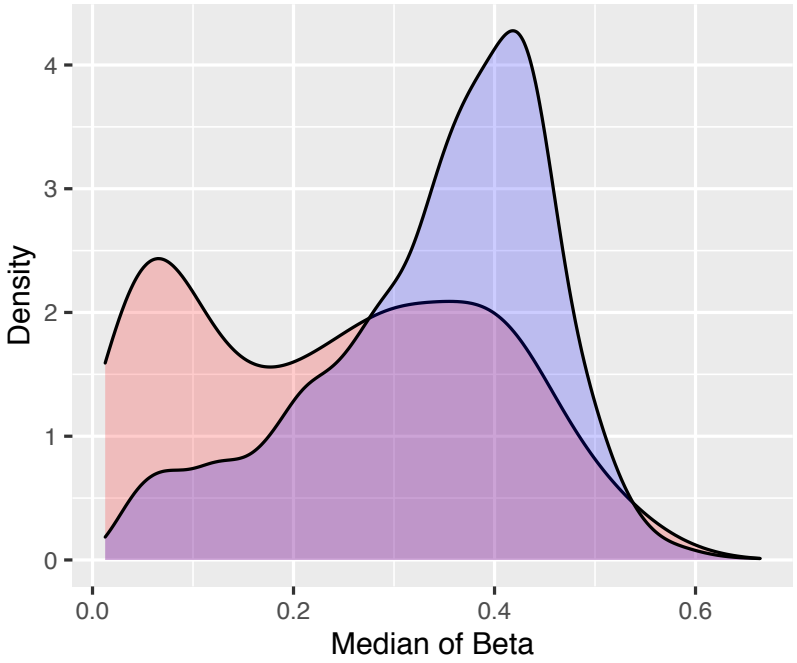


Figure B.2: The distribution of the median β value for probes targeting XCI escapees (red, 235 probes) or non-escapees (blue, 4050 probes) on the X chromosome of 273 female h-iPSCs. Genes which escaped XCI in all tissues in Tukiainen et al. 2017 are used as XCI escapees in this figure ($n = 99$). Among 235 probes targeting XCI escapee genes, 49% of probes display relatively high methylation level ($\beta > 0.25$) while 51% of probes display low methylation level ($\beta \leqslant 0.25$), indicating that in h-iPSCs, XCI might also happens on genes where were identified as escapees in other human tissues.

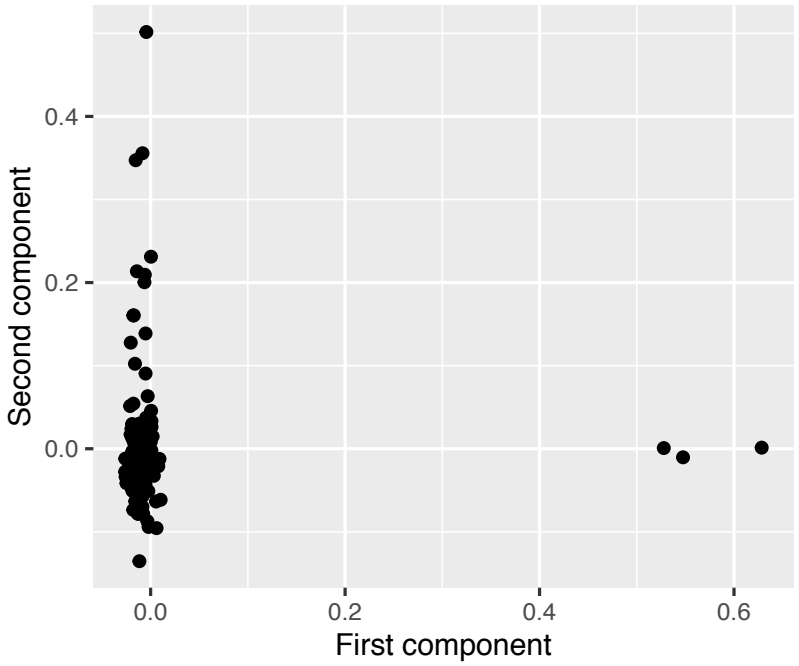


Figure B.3: Using PCA to investigate the population structure for 166 female h-iPSCs with pruned and filtered genetic variants: two groups are observed, whereas one group contains the vast majority of h-iPSCs (163 out of 166 h-iPSCs).

Bibliography

[10015] 1000 Genomes Project Consortium et al. "A global reference for human genetic variation". In: *Nature* 526.7571 (2015), pp. 68–74.

[Aby12] Alexej Abyzov et al. "Somatic copy number mosaicism in human skin revealed by induced pluripotent stem cells". In: *Nature* 492.7429 (2012), pp. 438–442.

[Ack18] Mania Ackermann et al. "Bioreactor-based mass production of human iPSC-derived macrophages enables immunotherapies against bacterial airway infections". In: *Nature communications* 9.1 (2018), pp. 1–13.

[Ala18] Kaur Alasoo et al. "Shared genetic effects on chromatin and gene expression indicate a role for enhancer priming in immune response". In: *Nature genetics* 50.3 (2018), pp. 424–431.

[Amp11] Katherine Amps et al. "Screening ethnically diverse human embryonic stem cells identifies a chromosome 20 minimal amplicon conferring growth advantage". In: *Nature biotechnology* 29.12 (2011), p. 1132.

[An12] Mahru C An et al. "Genetic correction of Huntington's disease phenotypes in induced pluripotent stem cells". In: *Cell stem cell* 11.2 (2012), pp. 253–263.

[Ana11] Gene Ananiev et al. "Isogenic pairs of wild type and mutant induced pluripotent stem cell (iPSC) lines from Rett syndrome patients as in vitro disease model". In: *PloS one* 6.9 (2011), e25255.

[APH15] Simon Anders, Paul Theodor Pyl, and Wolfgang Huber. "HTSeq—a Python framework to work with high-throughput sequencing data". In: *Bioinformatics* 31.2 (2015), pp. 166–169.

[AG84] Mary Anne Anderson and James F Gusella. "Use of cyclosporin A in establishing Epstein-Barr virus-transformed human lymphoblastoid cell lines". In: *In vitro* 20.11 (1984), pp. 856–858.

[Ang12] Montserrat C Anguera et al. "Molecular signatures of human induced pluripotent stem cells highlight sex differences and cancer genes". In: *Cell stem cell* 11.1 (2012), pp. 75–90.

[Ary14] Martin J Aryee et al. "Minfi: a flexible and comprehensive Bioconductor package for the analysis of Infinium DNA methylation microarrays". In: *Bioinformatics* 30.10 (2014), pp. 1363–1369.

[AH01] Philip Avner and Edith Heard. "X-chromosome inactivation: counting, choice and initiation". In: *Nature Reviews Genetics* 2.1 (2001), pp. 59–67.

[Bae86] Robert L Baehner et al. "DNA linkage analysis of X chromosome-linked chronic granulomatous disease". In: *Proceedings of the National Academy of Sciences* 83.10 (1986), pp. 3398–3401.

[BCB15] Bradley P Balaton, Allison M Cotton, and Carolyn J Brown. "Derivation of consensus inactivation status for X-linked genes from genome-wide studies". In: *Biology of sex differences* 6.1 (2015), pp. 1–11.

[Bal06] David J Balding. "A tutorial on statistical methods for population association studies". In: *Nature reviews genetics* 7.10 (2006), pp. 781–791.

[Ban18] Nicholas E Banovich et al. "Impact of regulatory variation across human iPSCs and differentiated cells". In: *Genome research* 28.1 (2018), pp. 122–131.

[Bar19] Shiran Bar et al. "Global characterization of X chromosome inactivation in human pluripotent stem cells". In: *Cell reports* 27.1 (2019), pp. 20–29.

[Bar15] Tahsin Stefan Barakat et al. "Stable X chromosome reactivation in female human induced pluripotent stem cells". In: *Stem cell reports* 4.2 (2015), pp. 199–208.

[Bat14] Douglas Bates et al. "Fitting linear mixed-effects models using lme4". In: *arXiv preprint arXiv:1406.5823* (2014).

[BH95] Yoav Benjamini and Yosef Hochberg. "Controlling the false discovery rate: a practical and powerful approach to multiple testing". In: *Journal of the Royal statistical society: series B (Methodological)* 57.1 (1995), pp. 289–300.

[BY01] Yoav Benjamini and Daniel Yekutieli. "The control of the false discovery rate in multiple testing under dependency". In: *Annals of statistics* (2001), pp. 1165–1188.

[Ber09] Ilse M van den Berg et al. "X chromosome inactivation is initiated in human preimplantation embryos". In: *The American Journal of Human Genetics* 84.6 (2009), pp. 771–779.

[Bia12a] Ilaria Bianchi et al. "The X chromosome and immune associated genes". In: *Journal of autoimmunity* 38.2-3 (2012), J187–J192.

[Bia12b] Ilaria Bianchi et al. "The X chromosome and immune associated genes". In: *Journal of autoimmunity* 38.2-3 (2012), J187–J192.

[BB12] Josipa Bilic and Juan Carlos Izpisua Belmonte. "Concise review: Induced pluripotent stem cells versus embryonic stem cells: close enough or yet too far apart?" In: *Stem cells* 30.1 (2012), pp. 33–41.

[BA86] J Martin Bland and DouglasG Altman. "Statistical methods for assessing agreement between two methods of clinical measurement". In: *The lancet* 327.8476 (1986), pp. 307–310.

[Bon36] Carlo Bonferroni. "Teoria statistica delle classi e calcolo delle probabilita". In: *Pubblicazioni del R Istituto Superiore di Scienze Economiche e Commericiali di Firenze* 8 (1936), pp. 3–62.

[Bra16] Nicolas L Bray et al. "Near-optimal probabilistic RNA-seq quantification". In: *Nature biotechnology* 34.5 (2016), pp. 525–527.

[Bre20] Alejandro J Brenes et al. "Erosion of human X chromosome inactivation causes major remodelling of the iPSC proteome". In: *bioRxiv* (2020).

[Bri15] Sharon F Briggs et al. "Single-Cell XIST Expression in Human Preimplantation Embryos and Newly Reprogrammed Female Induced Pluripotent Stem Cells". In: *Stem Cells* 33.6 (2015), pp. 1771–1781.

[Bri05] Thomas Heiberg Brix et al. "High frequency of skewed X-chromosome inactivation in females with autoimmune thyroid disease: a possible explanation for the female predisposition to thyroid autoimmunity". In: *The Journal of Clinical Endocrinology & Metabolism* 90.11 (2005), pp. 5949–5953.

[BT15] Neil Brockdorff and Bryan M Turner. "Dosage compensation in mammals". In: *Cold Spring Harbor perspectives in biology* 7.3 (2015), a019406.

[Bro91] Carolyn J Brown, Andrea Ballabio, et al. "A gene from the region of the human X inactivation centre is expressed exclusively from the inactive X chromosome". In: *Nature* 349.6304 (1991), pp. 38–44.

[Bro92] Carolyn J Brown, Brian D Hendrich, et al. "The human XIST gene: analysis of a 17 kb inactive X-specific RNA that contains conserved repeats and is highly localized within the nucleus". In: *Cell* 71.3 (1992), pp. 527–542.

[BB93] RM Brown and GK Brown. "X chromosome inactivation and the diagnosis of X linked disease in females." In: *Journal of medical genetics* 30.3 (1993), pp. 177–184.

[BJM17] Julian Buchrieser, William James, and Michael D Moore. "Human induced pluripotent stem cell-derived macrophages share ontogeny with MYB-independent tissue-resident macrophages". In: *Stem cell reports* 8.2 (2017), pp. 334–345.

[CW05] Laura Carrel and Huntington F Willard. "X-inactivation profile reveals extensive variability in X-linked gene expression in females". In: *Nature* 434.7031 (2005), pp. 400–404.

[CTJ13] Christopher A Cassa, Mark Y Tong, and Daniel M Jordan. "Large numbers of genetic variants considered to be pathogenic are common in asymptomatic individuals". In: *Human mutation* 34.9 (2013), pp. 1216–1220.

[Cas11] Raphaële Castagné et al. "Influence of sex and genetic variability on expression of X-linked genes in human monocytes". In: *Genomics* 98.5 (2011), pp. 320–326.

[Cha03] Juliet M Chapman et al. "Detecting disease associations due to linkage disequilibrium using haplotype tags: a class of tests and the determinants of statistical power". In: *Human heredity* 56.1-3 (2003), pp. 18–31.

[Che11] Guokai Chen et al. "Chemically defined conditions for human iPSC derivation and culture". In: *Nature methods* 8.5 (2011), pp. 424–429.

[CS09] Vivian G Cheung and Richard S Spielman. "Genetics of human gene expression: mapping DNA variants that influence gene expression". In: *Nature Reviews Genetics* 10.9 (2009), pp. 595–604.

[Chi09] Mark H Chin et al. "Induced pluripotent stem cells and embryonic stem cells are distinguished by gene expression signatures". In: *Cell stem cell* 5.1 (2009), pp. 111–123.

[Cho09] Kyung-Dal Choi et al. "Hematopoietic and endothelial differentiation of human induced pluripotent stem cells". In: *Stem cells* 27.3 (2009), pp. 559–567.

[Cho08] Edwin Choy et al. "Genetic analysis of human traits in vitro: drug response and gene expression in lymphoblastoid cell lines". In: *PLoS Genet* 4.11 (2008), e1000287.

[Cle96] Christine Moulton Clemson et al. "XIST RNA paints the inactive X chromosome at interphase: evidence for a novel RNA involved in nuclear/chromosome structure." In: *The Journal of cell biology* 132.3 (1996), pp. 259–275.

[Con05] International HapMap Consortium et al. "A haplotype map of the human genome". In: *Nature* 437.7063 (2005), p. 1299.

[Cos00] Carl Costanzi et al. "Histone macroH2A1 is concentrated in the inactive X chromosome of female preimplantation mouse embryos". In: *Development* 127.11 (2000), pp. 2283–2289.

[Cot15] Allison M Cotton et al. "Landscape of DNA methylation on the X chromosome reflects CpG density, functional chromatin state and X-chromosome inactivation". In: *Human molecular genetics* 24.6 (2015), pp. 1528–1539.

[DAn19] Agnieszka D'Antonio-Chronowska et al. "Association of human iPSC gene signatures and X chromosome dosage with two distinct cardiac differentiation trajectories". In: *Stem cell reports* 13.5 (2019), pp. 924–938.

[DSW10] Alan Dabney, John D Storey, and GR Warnes. "qvalue: Q-value estimation for false discovery rate control". In: *R package version* 1.0 (2010).

[Dan11] Petr Danecek, Adam Auton, et al. "The variant call format and VCFtools". In: *Bioinformatics* 27.15 (2011), pp. 2156–2158.

[Dan16] Petr Danecek, Shane A McCarthy, et al. "A method for checking genomic integrity in cultured cell lines from SNP genotyping data". In: *PLoS One* 11.5 (2016), e0155014.

[DSD14] Petr Danecek, Stephan Schiffels, and Richard Durbin. *Multiallelic calling model in bcftools (-m)*. 2014.

[De 06] Cosimo De Bari et al. "Mesenchymal multipotency of adult human periosteal cells demonstrated by single-cell lineage analysis". In: *Arthritis & Rheumatism* 54.4 (2006), pp. 1209–1221.

[DeB17] Christopher DeBoever et al. "Large-scale profiling reveals the influence of genetic variation on gene expression in human induced pluripotent stem cells". In: *Cell stem cell* 20.4 (2017), pp. 533–546.

[DMZ12] Olivier Delaneau, Jonathan Marchini, and Jean-François Zagury. "A linear complexity phasing method for thousands of genomes". In: *Nature methods* 9.2 (2012), pp. 179–181.

[Dob13] Alexander Dobin et al. "STAR: ultrafast universal RNA-seq aligner". In: *Bioinformatics* 29.1 (2013), pp. 15–21.

[Dor13] Carsten F Dormann et al. "Collinearity: a review of methods to deal with it and a simulation study evaluating their performance". In: *Ecography* 36.1 (2013), pp. 27–46.

[DS19] Michael Xavier Doss and Agapios Sachinidis. "Current challenges of iPSC-based disease modeling and therapeutic implications". In: *Cells* 8.5 (2019), p. 403.

[DC14] Jennifer A Doudna and Emmanuelle Charpentier. "The new frontier of genome engineering with CRISPR-Cas9". In: *Science* 346.6213 (2014).

[Dun18] Christopher G Duncan et al. "Dosage compensation and DNA methylation landscape of the X chromosome in mouse liver". In: *Scientific reports* 8.1 (2018), pp. 1–17.

[Dun61] Olive Jean Dunn. "Multiple comparisons among means". In: *Journal of the American statistical association* 56.293 (1961), pp. 52–64.

[ENS] ENSEMBL. *ENSEMBL version 75 human genome*. URL: https://grch37.ensembl.org/index.html (visited on 09/30/2010).

[Erh03] Sylvia Erhardt et al. "Consequences of the depletion of zygotic and embryonic enhancer of zeste 2 during preimplantation mouse development". In: *Development* 130.18 (2003), pp. 4235–4248.

[Est10] Miguel Angel Esteban et al. "Vitamin C enhances the generation of mouse and human induced pluripotent stem cells". In: *Cell stem cell* 6.1 (2010), pp. 71–79.

[Fad16] João Fadista et al. “The (in) famous GWAS P-value threshold revisited and updated for low-frequency variants”. In: *European Journal of Human Genetics* 24.8 (2016), pp. 1202–1205.

[FT11] Guoping Fan and Jamie Tran. “X chromosome inactivation in human and mouse pluripotent stem cells”. In: *Human genetics* 130.2 (2011), pp. 217–222.

[FL10] Jianqing Fan and Jinchi Lv. “A selective overview of variable selection in high dimensional feature space”. In: *Statistica Sinica* 20.1 (2010), p. 101.

[Fre11] Matthew L Freedman et al. “Principles for the post-GWAS functional characterization of cancer risk loci”. In: *Nature genetics* 43.6 (2011), p. 513.

[FHT10] Jerome Friedman, Trevor Hastie, and Rob Tibshirani. “Regularization paths for generalized linear models via coordinate descent”. In: *Journal of statistical software* 33.1 (2010), p. 1.

[GH18] Rafael Galupa and Edith Heard. “X-chromosome inactivation: a crossroads between chromosome architecture and gene regulation”. In: *Annual review of genetics* 52 (2018), pp. 535–566.

[GKQ14] Charles Gawad, Winston Koh, and Stephen R Quake. “Dissecting the clonal origins of childhood acute lymphoblastic leukemia by single-cell genomics”. In: *Proceedings of the National Academy of Sciences* 111.50 (2014), pp. 17947–17952.

[GKQ16] Charles Gawad, Winston Koh, and Stephen R Quake. “Single-cell genome sequencing: current state of the science”. In: *Nature Reviews Genetics* 17.3 (2016), p. 175.

[Gee16] M Geens et al. “Female human pluripotent stem cells rapidly lose X chromosome inactivation marks and progress to a skewed methylation pattern during culture”. In: *Molecular human reproduction* 22.4 (2016), pp. 285–298.

[Gor03] Siamon Gordon. “Alternative activation of macrophages”. In: *Nature reviews immunology* 3.1 (2003), pp. 23–35.

[Gor11] Athurva Gore et al. “Somatic coding mutations in human induced pluripotent stem cells”. In: *Nature* 471.7336 (2011), pp. 63–67.

[Han16] Hansen. “IlluminaHumanMethylation450kanno. ilmn12. hg19: annotation for illumina’s 450k methylation arrays”. In: *R package version 0.2* 1 (2016).

[Har12] Jennifer Harrow et al. “GENCODE: the reference human genome annotation for The ENCODE Project”. In: *Genome research* 22.9 (2012), pp. 1760–1774.

[HRW03] Cristina Hartshorn, John E Rice, and Lawrence J Wangh. "Differential pattern of Xist RNA accumulation in single blastomeres isolated from 8-cell stage mouse embryos following laser zona drilling". In: *Molecular Reproduction and Development: Incorporating Gamete Research* 64.1 (2003), pp. 41–51.

[HTF09] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The elements of statistical learning: data mining, inference, and prediction*. Springer Science & Business Media, 2009.

[HCA97] Edith Heard, Philippe Clerc, and Philip Avner. "X-chromosome inactivation in mammals". In: *Annual review of genetics* 31.1 (1997), pp. 571–610.

[HM07] A Helena Mangs and Brian J Morris. "The human pseudoautosomal region (PAR): origin, function and future". In: *Current genomics* 8.2 (2007), pp. 129–136.

[HC07] Asaf Hellman and Andrew Chess. "Gene body-specific methylation on the active X chromosome". In: *science* 315.5815 (2007), pp. 1141–1143.

[HR68] WG Hill and Alan Robertson. "Linkage disequilibrium in finite populations". In: *Theoretical and applied genetics* 38.6 (1968), pp. 226–231.

[HS16] Gabriel E Hoffman and Eric E Schadt. "variancePartition: interpreting drivers of variation in complex gene expression studies". In: *BMC bioinformatics* 17.1 (2016), pp. 1–13.

[Hol79] Sture Holm. "A simple sequentially rejective multiple test procedure". In: *Scandinavian journal of statistics* (1979), pp. 65–70.

[HDM09] Bryan Howie, Peter Donnelly, and Jonathan Marchini. "A flexible and accurate genotype imputation method for the next generation of genome-wide association studies". In: *PLoS Genet* 5.6 (2009), e1000529.

[Hu10a] Bao-Yang Hu et al. "Neural differentiation of human induced pluripotent stem cells follows developmental principles but with variable potency". In: *Proceedings of the National Academy of Sciences* 107.9 (2010), pp. 4335–4340.

[Hu10b] Bao-Yang Hu et al. "Neural differentiation of human induced pluripotent stem cells follows developmental principles but with variable potency". In: *Proceedings of the National Academy of Sciences* 107.9 (2010), pp. 4335–4340.

[Hub02] Wolfgang Huber et al. "Variance stabilization applied to microarray data calibration and to the quantification of differential expression". In: *Bioinformatics* 18.suppl_1 (2002), S96–S104.

[HL03] Khanh D Huynh and Jeannie T Lee. "Inheritance of a pre-inactivated paternal X chromosome in early mouse embryos". In: *Nature* 426.6968 (2003), pp. 857–862.

[Ima12] Yoichi Imaizumi et al. "Mitochondrial dysfunction associated with increased oxidative stress and α-synuclein accumulation in PARK2 iPSC-derived neurons and postmortem brain tissue". In: *Molecular brain* 5.1 (2012), pp. 1–13.

[Jam13] Gareth James et al. *An introduction to statistical learning*. Vol. 112. Springer, 2013.

[Jan19] Adrian Janiszewski et al. "Dynamic reversal of random X-Chromosome inactivation during iPSC reprogramming". In: *Genome research* 29.10 (2019), pp. 1659–1672.

[Jed18] Max Kuhn. Contributions from Jed Wing et al. *caret: Classification and Regression Training*. R package version 6.0-81. 2018. URL: https://CRAN.R-project.org/package=caret.

[JM04] Martin H Johnson and Josie ML McConnell. "Lineage allocation and cell polarity during mouse embryogenesis". In: *Seminars in cell & developmental biology*. Vol. 15. 5. Elsevier. 2004, pp. 583–597.

[Jor99] Michael I Jordan et al. "An introduction to variational methods for graphical models". In: *Machine learning* 37.2 (1999), pp. 183–233.

[KQ11] Tomer Kalisky and Stephen R Quake. "Single-cell genomics". In: *Nature methods* 8.4 (2011), pp. 311–314.

[Kan10] Hyun Min Kang et al. "Variance component model to account for sample structure in genome-wide association studies". In: *Nature genetics* 42.4 (2010), pp. 348–354.

[Ken14] Phillip J Kenny et al. "MOV10 and FMRP regulate AGO2 association with microRNA recognition elements". In: *Cell reports* 9.5 (2014), pp. 1729–1741.

[Kil17] Helena Kilpinen et al. "Common genetic variation drives molecular heterogeneity in human iPSCs". In: *Nature* 546.7658 (2017), pp. 370–375.

[KHP11] Kun-Yong Kim, Eriona Hysolli, and In-Hyun Park. "Neuronal maturation defect in induced pluripotent stem cells from patients with Rett syndrome". In: *Proceedings of the National Academy of Sciences* 108.34 (2011), pp. 14169–14174.

[Kim14] Kun-Yong Kim, Eriona Hysolli, Yoshiaki Tanaka, et al. "X chromosome of female cells shows dynamic changes in status during human somatic cell reprogrammingtrokovic2015combined". In: *Stem cell reports* 2.6 (2014), pp. 896–909.

[Kno09] Paul S Knoepfler. "Deconstructing stem cell tumorigenicity: a roadmap to safe regenerative medicine". In: *Stem cells* 27.5 (2009), pp. 1050–1056.

[KF13] Arthur Korte and Ashley Farlow. "The advantages and limitations of trait analysis with GWAS: a review". In: *Plant methods* 9.1 (2013), pp. 1–9.

[Kyu10] Minjung Kyung et al. "Penalized regression, standard errors, and Bayesian lassos". In: *Bayesian Analysis* 5.2 (2010), pp. 369–411.

[Lau11] Louise C Laurent et al. "Dynamic changes in the copy number of pluripotency and cell proliferation genes in human ESCs and iPSCs during reprogramming and time in culture". In: *Cell stem cell* 8.1 (2011), pp. 106–118.

[LKG18] Christopher ZW Lee, Tatsuya Kozaki, and Florent Ginhoux. "Studying tissue macrophages in vitro: are iPSC-derived cells the answer?" In: *Nature Reviews Immunology* 18.11 (2018), pp. 716–725.

[LB13] Jeannie T Lee and Marisa S Bartolomei. "X-inactivation, imprinting, and long noncoding RNAs in health and disease". In: *Cell* 152.6 (2013), pp. 1308–1323.

[LJ97] Jeannie T Lee and Rudolf Jaenisch. "Long-range cis effects of ectopic X-inactivation centres on a mouse autosome". In: *Nature* 386.6622 (1997), pp. 275–279.

[LDW99] Jeannie Lee, Lance S Davidow, and David Warshawsky. "Tsix, a gene antisense to Xist at the X-inactivation centre". In: *Nature genetics* 21.4 (1999), pp. 400–404.

[LK60] RC Lewontin and Ken-ichi Kojima. "The evolutionary dynamics of complex polymorphisms". In: *Evolution* 14.4 (1960), pp. 458–472.

[Li11] Heng Li. "A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data". In: *Bioinformatics* 27.21 (2011), pp. 2987–2993.

[LD09] Heng Li and Richard Durbin. "Fast and accurate short read alignment with Burrows–Wheeler transform". In: *bioinformatics* 25.14 (2009), pp. 1754–1760.

[Li09] Heng Li, Bob Handsaker, et al. "The sequence alignment/map format and SAMtools". In: *Bioinformatics* 25.16 (2009), pp. 2078–2079.

[Li18] Ye Li et al. "Human iPSC-derived natural killer cells engineered with chimeric antigen receptors enhance anti-tumor activity". In: *Cell stem cell* 23.2 (2018), pp. 181–192.

[LZ13] Gaoyang Liang and Yi Zhang. "Genetic and epigenetic variations in iPSCs: potential causes and implications for application". In: *Cell stem cell* 13.2 (2013), pp. 149–159.

[LDP10] Claude Libert, Lien Dejager, and Iris Pinheiro. "The X chromosome in immune functions: when a chromosome makes the difference". In: *Nature Reviews Immunology* 10.8 (2010), pp. 594–604.

[Lin97] Xihong Lin. "Variance component testing in generalised linear models with random effects". In: *Biometrika* 84.2 (1997), pp. 309–326.

[Lin18] Yuan-Ta Lin et al. "APOE4 causes widespread molecular and cellular alterations associated with Alzheimer's disease phenotypes in human iPSC-derived brain cell types". In: *Neuron* 98.6 (2018), pp. 1141–1154.

[Lin19] Stephanie M Linker et al. "Combined single-cell profiling of expression and DNA methylation reveals splicing regulation and heterogeneity". In: *Genome biology* 20.1 (2019), p. 30.

[Llo82] Stuart Lloyd. "Least squares quantization in PCM". In: *IEEE transactions on information theory* 28.2 (1982), pp. 129–137.

[LHA14] Michael I Love, Wolfgang Huber, and Simon Anders. "Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2". In: *Genome biology* 15.12 (2014), p. 550.

[Lud98] John Ludbrook. "Multiple comparison procedures updated". In: *Clinical and Experimental Pharmacology and Physiology* 25.12 (1998), pp. 1032–1037.

[Lui18] René Luijk et al. "Autosomal genetic variation is associated with DNA methylation in regions variably escaping X-chromosome inactivation". In: *Nature communications* 9.1 (2018), pp. 1–9.

[Lyo61] Mary F Lyon. "Gene action in the X-chromosome of the mouse (Mus musculus L.)" In: *nature* 190.4773 (1961), pp. 372–373.

[Mac18] Lantz C Mackey et al. "Epigenetic enzymes, age, and ancestry regulate the efficiency of human iPSC reprogramming". In: *Stem Cells* 36.11 (2018), pp. 1697–1708.

[Mac67] James MacQueen et al. "Some methods for classification and analysis of multivariate observations". In: *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*. Vol. 1. 14. Oakland, CA, USA. 1967, pp. 281–297.

[MHG16] Cheryl Maduro, Bas de Hoon, and Joost Gribnau. "Fitting the puzzle pieces: the bigger picture of XCI". In: *Trends in Biochemical Sciences* 41.2 (2016), pp. 138–147.

[MB12] Salah Mahmoudi and Anne Brunet. "Aging and reprogramming: a two-way street". In: *Current opinion in cell biology* 24.6 (2012), pp. 744–756.

[Mar10] Maria CN Marchetto et al. "A model for neural development and treatment of Rett syndrome using human induced pluripotent stem cells". In: *Cell* 143.4 (2010), pp. 527–539.

[Mar08] Fernando Oneissi Martinez et al. "Macrophage activation and polarization." In: *Frontiers in bioscience: a journal and virtual library* 13 (2008), p. 453.

[Mar11] Kristen Martins-Taylor, Benjamin S Nisler, et al. "Recurrent copy number variations in human induced pluripotent stem cells". In: *Nature biotechnology* 29.6 (2011), pp. 488–491.

[MX12] Kristen Martins-Taylor and Ren-He Xu. "Concise review: genomic stability of human induced pluripotent stem cells". In: *Stem Cells* 30.1 (2012), pp. 22–27.

[Mek12] Shila Mekhoubad et al. "Erosion of dosage compensation impacts human iPSC disease modeling". In: *Cell stem cell* 10.5 (2012), pp. 595–609.

[Mer18] Jerome Mertens et al. "Aging in a dish: iPSC-derived and directly induced neurons for studying brain aging and age-related neurodegenerative diseases". In: *Annual review of genetics* 52 (2018), pp. 271–293.

[Mon04] SA Monks et al. "Genetic inheritance of gene expression in human cell lines". In: *The American Journal of Human Genetics* 75.6 (2004), pp. 1094–1105.

[MLF13] Lisa D Moore, Thuc Le, and Guoping Fan. "DNA methylation and its basic function". In: *Neuropsychopharmacology* 38.1 (2013), pp. 23–38.

[Mül11] Franz-Josef Müller et al. "A bioinformatic assay for pluripotency in human cells". In: *Nature methods* 8.4 (2011), pp. 315–317.

[Nav11] Nicholas Navin et al. "Tumour evolution inferred by single-cell sequencing". In: *Nature* 472.7341 (2011), p. 90.

[Naz12] Kristopher L Nazor et al. "Recurrent variations in DNA methylation in human pluripotent stem cells and their differentiated derivatives". In: *Cell stem cell* 10.5 (2012), pp. 620–634.

[Nei86] Heidemarie Neitzel. "A routine method for the establishment of permanent growing lymphoblastoid cell lines". In: *Human genetics* 73.4 (1986), pp. 320–326.

[NC10] Aaron M Newman and James B Cooper. "Lab-specific gene expression signatures in pluripotent stem cells". In: *Cell stem cell* 7.2 (2010), pp. 258–262.

[Niu10] Nifang Niu et al. "Radiation pharmacogenomics: a genome-wide association approach to identify radiation response biomarkers using human lymphoblastoid cell lines". In: *Genome research* 20.11 (2010), pp. 1482–1492.

[Oka04] Ken-Ichi Okamoto et al. "Rapid and persistent modulation of actin dynamics regulates postsynaptic reorganization underlying bidirectional plasticity". In: *Nature neuroscience* 7.10 (2004), pp. 1104–1112.

[PPG15] Athma A Pai, Jonathan K Pritchard, and Yoav Gilad. "The genetic and mechanistic basis for variation in gene regulation". In: *PLoS Genet* 11.1 (2015), e1004857.

[PJ96] Barbara Panning and Rudolf Jaenisch. "DNA hypomethylation can activate Xist expression and silence X-linked genes." In: *Genes & development* 10.16 (1996), pp. 1991–2002.

[Pan17] Athanasia D Panopoulos et al. "iPSCORE: a resource of 222 iPSC lines enabling functional characterization of genetic variation across a variety of cell types". In: *Stem cell reports* 8.4 (2017), pp. 1086–1100.

[Paş11] Sergiu P Paşca et al. "Using iPSC-derived neurons to uncover cellular phenotypes associated with Timothy syndrome". In: *Nature medicine* 17.12 (2011), p. 1657.

[PP15] Vincent Pasque and Kathrin Plath. "X chromosome reactivation in reprogramming and in development". In: *Current opinion in cell biology* 37 (2015), pp. 75–83.

[Pen96] Graeme D Penny et al. "Requirement for Xist in X chromosome inactivation". In: *Nature* 379.6561 (1996), pp. 131–137.

[Pha13] Paul DP Pharoah et al. "GWAS meta-analysis and replication identifies three new susceptibility loci for ovarian cancer". In: *Nature genetics* 45.4 (2013), pp. 362–370.

[Phi08] Theresa Phillips et al. "The role of methylation in gene expression". In: *Nature Education* 1.1 (2008), p. 116.

[Pic14] Simone Picelli et al. "Full-length RNA-seq from single cells using Smart-seq2". In: *Nature protocols* 9.1 (2014), pp. 171–181.

[PLI] PLINK. *PLINK version 1.9*. URL: http://zzz.bwh.harvard.edu/plink/data.shtml (visited on 09/30/2010).

[Pom11] Oz Pomp et al. "Unexpected X chromosome skewing during culture and reprogramming of human somatic cells can be alleviated by exogenous telomerase". In: *Cell stem cell* 9.2 (2011), pp. 156–165.

[PG11] Daphne B Pontier and Joost Gribnau. "Xist regulation and function explored". In: *Human genetics* 130.2 (2011), pp. 223–236.

[Pur07] Shaun Purcell et al. "PLINK: a tool set for whole-genome association and population-based linkage analyses". In: *The American journal of human genetics* 81.3 (2007), pp. 559–575.

[Qui13] Emma M Quinn et al. "Development of strategies for SNP detection in RNA-seq data: application to lymphoblastoid cell lines and evaluation using 1000 Genomes data". In: *PloS one* 8.3 (2013), e58815.

[R C17] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria, 2017. URL: https://www.R-project.org/.

[Raj11] Deepika Rajesh et al. "Human lymphoblastoid B-cell lines reprogrammed to EBV-free induced pluripotent stem cells". In: *Blood* 118.7 (2011), pp. 1797–1800.

[Ran13] F Ann Ran et al. "Genome engineering using the CRISPR-Cas9 system". In: *Nature protocols* 8.11 (2013), pp. 2281–2308.

[RM12] Mahendra S Rao and Nasir Malik. "Assessing iPSC reprogramming methods for their suitability in translational medicine". In: *Journal of cellular biochemistry* 113.10 (2012), pp. 3061–3068.

[RC15] Terje Raudsepp and Bhanu P Chowdhary. "The eutherian pseudoautosomal region". In: *Cytogenetic and Genome Research* 147.2-3 (2015), pp. 81–94.

[Roy82a] J Patrick Royston. "An extension of Shapiro and Wilk's W test for normality to large samples". In: *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 31.2 (1982), pp. 115–124.

[Roy82b] JP Royston. "Algorithm AS 181: the W test for normality". In: *Journal of the Royal Statistical Society. Series C (Applied Statistics)* 31.2 (1982), pp. 176–180.

[Roy95] Patrick Royston. "Remark AS R94: A remark on algorithm AS 181: The W-test for normality". In: *Journal of the Royal Statistical Society. Series C (Applied Statistics)* 44.4 (1995), pp. 547–551.

[San18] Suttikarn Santiwatana et al. "Skewed X chromosome inactivation in girls and female adolescents with autoimmune thyroid disease". In: *Clinical Endocrinology* 89.6 (2018), pp. 863–869.

[Sar17] Valentina Lo Sardo et al. "Influence of donor age on induced pluripotent stem cells". In: *Nature biotechnology* 35.1 (2017), pp. 69–74.

[Sch14] David C Schöndorf et al. "iPSC-derived neurons from GBA1-associated Parkinson's disease patients show autophagic defects and impaired calcium homeostasis". In: *Nature communications* 5.1 (2014), pp. 1–17.

[Sch18] Jeremy Schwartzentruber et al. "Molecular and functional variation in iPSC-derived sensory neurons". In: *Nature genetics* 50.1 (2018), pp. 54–61.

[Šid67] Zbyněk Šidák. "Rectangular confidence regions for the means of multivariate normal distributions". In: *Journal of the American Statistical Association* 62.318 (1967), pp. 626–633.

[Sim13] Matthew D Simon et al. "High-resolution Xist binding maps reveal two-step spreading during X-chromosome inactivation". In: *Nature* 504.7480 (2013), pp. 465–469.

[Sin15] Vimal K Singh et al. "Induced pluripotent stem cells: applications in regenerative medicine, disease modeling, and drug discovery". In: *Frontiers in cell and developmental biology* 3 (2015), p. 2.

[Sla08] Montgomery Slatkin. "Linkage disequilibrium—understanding the evolutionary past and mapping the medical future". In: *Nature Reviews Genetics* 9.6 (2008), pp. 477–485.

[Spi08] Claudia Spits et al. "Recurrent chromosomal abnormalities in human embryonic stem cells". In: *Nature biotechnology* 26.12 (2008), pp. 1361–1363.

[Spl11] Erik Splinter et al. "The inactive X chromosome adopts a unique three-dimensional conformation that is dependent on Xist RNA". In: *Genes & development* 25.13 (2011), pp. 1371–1383.

[Ste72] CM Steel. "Human lymphoblastoid cell lines. III. Co-cultivation technique for establishment of new lines". In: *Journal of the National Cancer Institute* 48.3 (1972), pp. 623–628.

[Ste10] Oliver Stegle, Leopold Parts, Richard Durbin, et al. "A Bayesian framework to account for complex non-genetic factors in gene expression levels greatly increases power in eQTL studies". In: *PLoS computational biology* 6.5 (2010).

[Ste12] Oliver Stegle, Leopold Parts, Matias Piipari, et al. "Using probabilistic estimation of expression residuals (PEER) to obtain increased power and interpretability of gene expression analyses". In: *Nature protocols* 7.3 (2012), p. 500.

[Sud15] Peter H Sudmant et al. "An integrated map of structural variation in 2,504 human genomes". In: *Nature* 526.7571 (2015), pp. 75–81.

[Taa11] Seth M Taapken et al. "Karyotypic abnormalities in human induced pluripotent stem cells and embryonic stem cells". In: *Nature biotechnology* 29.4 (2011), pp. 313–314.

[Tak07] Kazutoshi Takahashi, Koji Tanabe, et al. "Induction of pluripotent stem cells from adult human fibroblasts by defined factors". In: *cell* 131.5 (2007), pp. 861–872.

[TY06] Kazutoshi Takahashi and Shinya Yamanaka. "Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors". In: *cell* 126.4 (2006), pp. 663–676.

[TBT08] Zohreh Talebizadeh, Merlin G Butler, and Mariana F Theodoro. "Feasibility and relevance of examining lymphoblastoid cell lines to study role of microRNAs in autism". In: *Autism Research* 1.4 (2008), pp. 240–250.

[Tch10] Jason Tchieu et al. "Female human iPSCs retain an inactive X chromosome". In: *Cell stem cell* 7.3 (2010), pp. 329–342.

[Ten12] Jacob A Tennessen et al. "Evolution and functional impact of rare coding variation from deep sequencing of human exomes". In: *science* 337.6090 (2012), pp. 64–69.

[Tib96] Robert Tibshirani. "Regression shrinkage and selection via the lasso". In: *Journal of the Royal Statistical Society: Series B (Methodological)* 58.1 (1996), pp. 267–288.

[TVB11] Gustavo Tiscornia, Erica Lorenzo Vivas, and Juan Carlos Izpisúa Belmonte. "Diseases in a dish: modeling human genetic disorders using induced pluripotent cells". In: *Nature medicine* 17.12 (2011), pp. 1570–1576.

[Tom12] Kiichiro Tomoda et al. "Derivation conditions impact X-inactivation status in female human induced pluripotent stem cells". In: *Cell stem cell* 11.1 (2012), pp. 91–99.

[Tri92] Carla Tribioli et al. "Methylation and sequence analysis around EagI sites: identification of 28 new CpG islands in XQ24-XQ28". In: *Nucleic acids research* 20.4 (1992), pp. 727–733.

[Tro15] Ras Trokovic et al. "Combined negative effect of donor age and time in culture on the reprogramming efficiency into induced pluripotent stem cells". In: *Stem Cell Research* 15.1 (2015), pp. 254–262.

[Tsu19] Osahiko Tsuji et al. "Concise review: laying the groundwork for a first-in-human study of an induced pluripotent stem cell-based intervention for spinal cord injury". In: *Stem Cells* 37.1 (2019), pp. 6–13.

[Tuk17] Taru Tukiainen et al. "Landscape of X chromosome inactivation across human tissues". In: *Nature* 550.7675 (2017), pp. 244–248.

[Vac16] Marcella Vacca et al. "X inactivation and reactivation in X-linked diseases". In: *Seminars in cell & developmental biology*. Vol. 56. Elsevier. 2016, pp. 78–87.

[Val13] Céline Vallot, Christophe Huret, et al. "XACT, a long noncoding transcript coating the active X chromosome in human pluripotent cells". In: *Nature genetics* 45.3 (2013), pp. 239–241.

[Val15] Céline Vallot, Jean-François Ouimette, et al. "Erosion of X chromosome inactivation in human pluripotent cells initiates with XACT coating and depends on a specific heterochromatin landscape". In: *Cell stem cell* 16.5 (2015), pp. 533–546.

[Val17] Céline Vallot, Catherine Patrat, et al. "XACT noncoding RNA competes with XIST in the control of X chromosome activity during human early development". In: *Cell stem cell* 20.1 (2017), pp. 102–111.

[Van11] IM Van den Berg et al. "XCI in preimplantation mouse and human embryos: first there is remodelling..." In: *Human genetics* 130.2 (2011), pp. 203–215.

[Via20] Evelyn Quintanilha Vianna et al. "Understanding the Landscape of X-linked Variants Causing Intellectual Disability in Females Through Extreme X Chromosome Inactivation Skewing". In: *Molecular neurobiology* 57.9 (2020), pp. 3671–3684.

[Vol18] Viola Volpato et al. "Reproducibility of molecular phenotypes after long-term differentiation to human iPSC-derived neurons: a multisite omics study". In: *Stem cell reports* 11.4 (2018), pp. 897–911.

[Wan17] Gang Wang et al. "Efficient, footprint-free human iPSC genome editing by consolidation of Cas9/CRISPR and piggyBac technologies". In: *Nature protocols* 12.1 (2017), p. 88.

[Wan13] Su Wang et al. "Human iPSC-derived oligodendrocyte progenitor cells can myelinate and rescue a mouse model of congenital hypomyelination". In: *Cell stem cell* 12.2 (2013), pp. 252–264.

[WD12] Heather E Wheeler and M Eileen Dolan. "Lymphoblastoid cell lines in pharmacogenomic discovery and clinical translation". In: *Pharmacogenomics* 13.1 (2012), pp. 55–70.

[Wic16] Hadley Wickham. *ggplot2: elegant graphics for data analysis*. springer, 2016.

[Wu14] Hao Wu et al. "Cellular resolution maps of X chromosome inactivation: implications for neural development, function, and disease". In: *Neuron* 81.1 (2014), pp. 103–119.

[WH11] Sean M Wu and Konrad Hochedlinger. "Harnessing the potential of induced pluripotent stem cells for regenerative medicine". In: *Nature cell biology* 13.5 (2011), pp. 497–505.

[WJ00] Anton Wutz and Rudolf Jaenisch. "A shift from reversible to irreversible X inactivation is triggered during ES cell differentiation". In: *Molecular cell* 5.4 (2000), pp. 695–705.

[Xu14] ChangJiang Xu et al. "Estimating genome-wide significance for whole-genome sequencing studies". In: *Genetic epidemiology* 38.4 (2014), pp. 281–290.

[Yan10] Jian Yang et al. "Common SNPs explain a large proportion of the heritability for human height". In: *Nature genetics* 42.7 (2010), p. 565.

[Zer02] Magdalena Zernicka-Goetz. "Patterning of the embryo: the first spatial decisions in the life of a mouse". In: *Development* 129.4 (2002), pp. 815–829.

[Zha15] Hanrui Zhang et al. "Functional analysis and transcriptomic profiling of iPSC-derived macrophages and their application in modeling Mendelian disease". In: *Circulation research* 117.1 (2015), pp. 17–28.

[Zha08] Yang Zhao et al. "Two supporting factors greatly improve the efficiency of human iPSC generation". In: *Cell stem cell* 3.5 (2008), pp. 475–479.