Transcriptomic approaches for investigating developmental lineage: exploiting the X-chromosome as a marker for lineage specification and quantifying the lineage fidelity of neural organoid systems

A dissertation presented by

Jonathan Werner

to the

School of Biological Sciences

in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

in

Biological Sciences

at

Cold Spring Harbor Laboratory

September 2023

Acknowledgements

It is an immense privilege to have been given the opportunity for graduate study at CSHL; one that I do not take lightly and view as arising largely from the efforts of those who, for whichever reason, decided I was worth a moment of their time and invested into my future. I stand where I am today due to the actions of a large collection of friends, family, and colleagues, to all of whom I sincerely say thank you.

I would like to thank my research advisor, Jesse Gillis, for his time and thoughts over the years; one of the most rewarding aspects of my time at CSHL has been the continual intellectual challenges I've faced working in your lab, not just in my own research but in thinking about many areas of science. My experience in the Gillis lab was greatly enhanced thanks to my fellow lab members, whom I thank for all the conversations over the years. I'd particularly like to acknowledge Sara Ballouz, Maggie Crow, and Stephan Fischer for their patience in being bothered with questions and their willingness to help me get started in the lab. I would also like to thank the members of my thesis committee, Camila dos Santos, Dan Levy, and Adam Siepel for their guidance and offer my thanks to Christine Disteche for agreeing to be my external committee member.

To the staff of the CSHL School of Biological Sciences, thank you for the support and flexibility in my non-traditional tenure at CSHL. The ability to move back home to Baltimore in my last year was incredibly helpful and I am extremely grateful. Thank you for also giving me the chance to be a CSHL graduate student, the support we receive from the school is unparalleled in my opinion and I am very grateful to have benefitted from the care and dedication you all bring to CSHL. And to the Class of 2018, you inspire me as friends and colleagues and I look forward to seeing all of your future success. Especially to Alexa, Amritha, Marie, and Mo, CSHL would have been a much duller experience without you all.

My journey through CSHL would not have been nearly as smooth or even possible without my time at UMBC as an undergraduate, chiefly due to the Meyerhoff Scholars program. I learned a great many valuable lessons and was presented numerous opportunities as a Meyerhoff Scholar that I wouldn’t have otherwise and would like to thank the program staff for their continued support. I credit the basis of my current research interests to the experience gained during my time in Rachel Brewster’s lab at UMBC. I would like to thank Rachel for providing my first research experience and fostering a deep curiosity about developmental biology. And to Eudorah Vital, sitting next to you during selection weekend was one of the best happenstances of my life, it is a joy to count you as one of my greatest friends.

To my partner, Colin, thank you for your seemingly immeasurable patience and kindness. I am continually striving to be the partner you deserve and am deeply grateful for what we’ve built together. And, perhaps more importantly, I’d like to thank my cat, Pickles. Because of you, I’ve been able to give up alarm clocks and a peaceful night’s sleep. Who knew such a small cat could make so much noise, all the time, at any given moment.

And finally, I’d like to thank all the experimentalists who make their data publicly available, without whom, none of my work presented here would be possible.

Table of Contents

List of Abbreviations

AUROC	Area under receiver operating curve
EN-TEx	Epigenome and Tissue Expression dataset
eQTL	Expression quantitative trait loci
GO	Gene Ontology
GTEx	Genotype-Tissue Expression dataset
PCA	Principal Component Analysis
RME	Random monoallelic expression
RNA-seq	RNA-sequencing
scRNA-seq	single-cell RNA-sequencing
SNPs	Single Nucleotide Polymorphisms
SRA	Sequencing Read Archive
XCE	X-chromosome controlling element
XCI	X-chromosome inactivation

List of Figures

1. Introduction

Complex multicellular organisms are composed of an immense diversity of cellular phenotypes distributed among cells sharing nearly identical genetic material. Exactly how such variation in cellular phenotypes arises from identical genomes is rooted in the developmental lineage relationships that link all cells. When daughter cells inherit a copy of the genome after cellular division, their functional genomic outputs can significantly diverge, both from each other and from the original ancestral genome, generating phenotypic diversity. C.H. Waddington formalized a mechanism linking the production of cellular diversity through developmental lineage with the term ‘epigenetics’ (Waddington 1942b), defined as effects inherited through cell divisions that direct the functional output of the genome (gene expression) without altering the DNA sequence. This progressive modulation of functional genomic output via developmental lineage defines the multitude of canalized routes a genome takes when transitioning from totipotency to differentiated phenotypes (Waddington 1942a), extracting vast cellular diversity from a single originating totipotent stem cell. Many investigative paths surround the study of developmental lineage and this thesis focuses on two distinct themes; 1.) exploiting a large-scale, developmentally early epigenetic event (X-Chromosome Inactivation, XCI) as a marker of developmental lineage to investigate early lineage specification events in mammals, and 2.) the venture of modeling developmental lineage *in vitro* (organoid systems). This thesis is structured around 3 specific research aims, each addressed with a separate manuscript:

- Aim 1: Characterize early lineage events in human development through assessments of cross-tissue variability in XCI
- Aim 2: Investigate stochastic and genetic factors contributing to population variability in XCI across mammalian species

- Aim 3: Quantify the fidelity of human neural organoid systems to primary neural tissue

The introduction begins with a description of XCI along with an assessment of the field's current understanding of the interplay between XCI and developmental lineage, outlining how cross-tissue and cross-species analyses of XCI variability can answer outstanding questions. This is followed by a description of organoid technologies with a focus on neural organoids and current challenges relating *in vivo* and *in vitro* biology.

1.1. X-chromosome inactivation

Mammalian females and males differ genetically via their sex chromosomes, where females carry 2 X-chromosomes and males carry an X- and a Y-chromosome (Barr and Bertram 1949; Barr and Carr 1960). The Y-chromosome is extremely gene poor compared to the X-chromosome, containing only ~70 protein-coding genes to the ~850 X-linked genes (O'Leary et al. 2016). This large genetic imbalance produces a gene dosage discrepancy across the sexes, where females have twice the number of X-linked genes compared to males, theoretically generating a double dose of X-linked gene expression. Yet, for the most part, mammalian female cells do not produce twice the amount of X-linked gene product compared to male cells (Heard and Disteche 2006). This is due to X-chromosome inactivation (XCI), an epigenetic event that silences transcription from a single X-allele in all female cells, resulting in equivalent X-linked gene dosage across the sexes (Lyon 1961; Dossin and Heard 2021).

X-chromosome inactivation describes the massive epigenetic reconfiguration of a single X-allele in female cells to silence transcription (Dossin and Heard 2021), orchestrated through the long non-coding RNA XIST (Brown et al. 1992). XIST operates in cis, physically coating its associated allele and triggering a cascade of epigenetic modifications that transform the allele into a highly condensed heterochromatic state, inhibiting gene expression (Fang et al. 2019; Dixon-McDougall and Brown 2022). XCI occurs developmentally early, typically in the first days/weeks of embryogenesis prior to implantation into the uterine lining (van den Berg et al. 2009), with the exact timing variable across species (Lyon 1972). XCI is also random and permanent (Evans et al. 1965; Wu et al. 2014). Each cell makes an independent choice for which allele to inactivate, with the inactivated allele inherited through cell divisions. This propagates the initial random choice of allelic inactivation down each cell's subsequent lineage. As a combination of the early, random, and permanent nature of XCI, all female mammals are mosaics for X-linked gene expression (Migeon 2013), an enduring phenotypic consequence connecting all cells of an individual back to one of the earliest embryonic developmental milestones.

An additional significant characteristic of XCI is that XCI ratios, the ratio of inactivated parental X-alleles, are highly variable within adult populations, ranging from balanced to completely skewed (Amos-Landgraf et al. 2006; Shvetsova et al. 2019). The XCI ratio of an individual becomes highly consequential in the presence of disease-causing X-linked genetic variants, where the allelic direction and magnitude of XCI can directly influence manifestation of disease phenotypes (Migeon 2013). As such, it is a long-standing scientific effort to understand how population variability in XCI arises, with contemporary opinion favoring either direct or indirect genetic mechanisms (Belmont 1996; Migeon 1998; Brown and Robinson 2000).

This is derived from several lines of evidence, that extreme XCI ratios are consistently associated with disease phenotypes of genetic X-linked disorders in heterozygous female carriers (Migeon 1971; Migeon et al. 1981; Devriendt et al. 1997;

Plenge et al. 2002) and that in mice, there is a strong genetic basis of variable XCI across laboratory strains (Simmler et al. 1993; Calaway et al. 2013; Sun et al. 2021). However, in the disease context, it remains difficult to prove whether disease variants generally cause skewed XCI or whether it is the combination of a disease variant and skewed XCI that result in disease phenotypes. When considering genetic mechanisms for variable XCI in mice, strong evidence for comparable mechanisms in human populations remains to be found (Peeters et al. 2016). Taken together, genetic explanations are poor general models for the observed population variability in XCI ratios and overlook the fundamental characteristics of XCI that naturally produce variability across individuals: the early, permanent, and random nature of allelic-inactivation during XCI.

This thesis outlines how our current understanding of XCI variability is derived from an extremely narrow biological consideration, typically limited to data from single lineages/tissues (whole blood) and few mammalian species (human and mouse). The work presented here broadly extends analysis of XCI variability across numerous lineages/tissues in humans and across diverse mammalian species, addressing long-standing questions surrounding XCI variability. In Aim 1, extensive cross-tissue analysis of XCI in humans is performed and, through the lens of XCI as a marker for developmental lineage, characteristics of lineage specification are derived both at a broad organismal level and in a lineage-specific manner. In Aim 2, analysis of population XCI ratio variability is extended to 9 mammalian species, evaluating conserved aspects of XCI and exploring both stochastic and genetic models for explaining observed population XCI variability.

1.2. XCI as a marker for developmental lineage

Quickly following the initial hypothesis of random inactivation put forth by Mary Lyon in 1961 (Lyon 1961), it was recognized that the permanence of variable X-linked allelic expression within an individual could be exploited to infer developmental lineage characteristics from adult data. The simplest instance of this is inferring whether a lineage had a single or multiple cell origin (Linder and Gartler 1965; Fialkow et al. 1967, 1971; Gartler et al. 1966). Lineages with a single cell origin would always have the same active parental X-allele, whereas both active parental alleles could be present for multi-cellular lineages. This approach was used effectively in the study of cellular origins for tumors (Linder and Gartler 1965; Fialkow et al. 1967, 1971; Gartler et al. 1966) and is largely responsible for our current understanding that many tumors are derived from single starting cells. The essential idea is associating variance of XCI ratios within a lineage to the number of cells that must have been selected earlier in development for that lineage.

Extending beyond the question of single or multi-cellular origins of a lineage, the variance of XCI ratios can be used to infer the exact number of cells that must have been present at the time of lineage specification (Gartler et al. 1969). Take a population of cells that has an equal mix of inactivated parental alleles, where a subset of these cells is selected for a specific lineage. If only a single cell was selected for the lineage, only a single active parental allele would be present within the lineage. If two cells were selected for the lineage, X-linked allelic expression can either be completely skewed with both cells carrying an active paternal or maternal X-allele, or completely balanced with the cells carrying different active X-alleles. The more cells that are sampled for the lineage, the less variable XCI ratios will be within that lineage. This relationship, perfectly described through binomial sampling distributions, has been used to estimate the size of progenitor pools for tissues for decades, typically limited to accessible tissues like blood and skin in humans (Gandini and Gartler 1969; Gartler et al. 1969). Originally, heterozygotes with a clear phenotypic distinction between the X-alleles (G6PD deficiency) were used to estimate the XCI ratio of individuals in the 1960’s. More recent molecular approaches have removed the requirement for observably distinct X-alleles, quantifying the XCI ratio of tissue samples through allele-specific methylation(Amos-Landgraf et al. 2006; Allen et al. 1992) or expression patterns (Shvetsova et al. 2019; Szelinger et al. 2014). Importantly, most studies in humans are limited to accessible tissues and the vast majority of analysis for variance in XCI ratios is derived from whole blood samples. Not only is variance in XCI ratios informative for individual lineages, but it also is informative for inferring relationships across lineages and has yet to fully be exploited for large-scale lineage analysis within human data.

Inferring relationships across lineages from XCI ratio variability comes down to assessing the degree of shared variance in XCI ratios (Gandini et al. 1968; Nesbitt 1971; Fialkow 1973; McMahon et al. 1983). If two lineages are highly correlated in their XCI ratios, it can be inferred they likely shared a progenitor pool after XCI. The XCI ratios of the individual lineages would be sampled from the initial stochastically determined XCI ratio at the time of inactivation, establishing dependence across the lineages for XCI variability. If two lineages are largely uncorrelated and never share XCI ratios, XCI must have occurred independently across the lineages, i.e., after their specification. The first instance of assessing cross-tissue variability in XCI ratios to infer lineage relationships of non-accessible tissues was reported in 1971 through the work of Muriel Nesbitt (Nesbitt 1971). The work relied on a cytologically visible translocation (Cattanach's translocation) on the mouse X-chromosome to distinguish cells that carried a normal or translocated inactivated X-chromosome, thus enabling XCI ratios to be computed. Co-variance in XCI ratios across 5 different tissues were computed revealing substantial shared variability in XCI ratios across all lineages. As only tissues from the ectodermal and mesodermal germ-layers were sampled (germ-layer specification is introduced more thoroughly in section 1.3.), the conclusion was reached that XCI precedes the separation of ectoderm and mesoderm tissues, in a pool of approximately 13-21 progenitors. While limited in scale by the technologies at the time, the underlying logic and approaches for interrogating lineage relationships through XCI ratio variability presented within this early work are applicable to any measure of XCI ratios. Excitingly, more modern molecular approaches for estimating XCI ratios bring dramatic increases in scalability (see section 1.4.). A key scientific advance presented within this thesis is coupling these foundational ideas of characterizing developmental lineage through XCI variability with modern scalable measures of XCI ratios, to comprehensively map organism-wide lineage relationships in humans and, more broadly, perform extensive cross-species comparisons of population XCI variability.

1.3. Unknowns of human development

A hugely consequential lineage decision that may overlap with the timing of XCI is germ-layer specification (Ghimire et al. 2021). The ectodermal, endodermal, and mesodermal germ layers are specified from the inner cell mass of the blastocyst and go on to form all the tissues of the eventual individual, distinct from the extra-embryonic lineages that go on to form placental tissues. If germ-layer specification occurs after XCI, it is expected that XCI ratios will be shared across the germ-layers and subsequently across all tissues. If any degree of tissue specification occurs before XCI, XCI ratios across those lineages specified will be independent and are not expected to be shared. Presently, current evidence indicates XCI is initiated early in human development (van den Berg et al. 2009; Petropoulos et al. 2016; Moreira de Mello et al. 2017) , but the exact overlap in the completion of XCI and early lineage specification events, specifically embryonic tissue specification, is unclear. This can be attributed to the difficulty in studying human embryos, both ethically and technically, and the restriction to accessible tissues in adults. As it currently stands, the vast majority of our understanding relating XCI, developmental lineage, and adult XCI variability in humans stems from a single lineage, whole blood, which may not be generalizable when taking into account the developmental context of XCI.

1.4. Utility of cross-tissue analysis of XCI ratios

Several studies assessing XCI of whole blood samples in humans established XCI ratios are highly variable across individuals and are largely responsible for our understanding that XCI varies generally within adult populations (Amos-Landgraf et al. 2006; Shvetsova et al. 2019). However, as outlined above, the temporal relationship between XCI and early lineage specification has significant impacts on the variability of XCI across lineages, where observations from a single lineage may not be generalizable. This is clinically relevant as it is common practice to quantify an individual's XCI ratio through a blood sample to assess one's risk for X-linked disorder phenotypes. Typical approaches for comparing XCI ratios across tissues in humans are usually limited in the number of tissues assessed (Bittel et al. 2008; Hoon et al. 2015). While these studies report correlated XCI ratios across tissues, indicating XCI ratios are shared across lineages to an extent, a comprehensive assessment across lineages with extensive sampling of all germ-layer lineages is lacking.

Cross-tissue analysis of XCI ratios, when sampling across the three germ-layers, stands to address fundamental questions as to the timing of XCI and how XCI ratios present across lineages, revealing characteristics of developmental lineage at multiple scales. Assessments of shared XCI variance can reveal the temporal ordering of XCI and germ-layer specification and assessments of lineage-specific variance can infer lineage-specific progenitor cell counts. Such broad cross-tissue data for humans is present within the Genotype-Tissue Expression dataset (Lonsdale et al. 2013), which samples bulk RNA-sequencing (RNA-seq) data from numerous tissues across hundreds of individuals. A key component of this thesis is the development of an approach to model XCI ratios from bulk RNA-seq data, enabling extensive cross-tissue analysis of XCI in humans and cross-species assessments.

1.5. Modeling XCI ratios from bulk RNA-sequencing data

Bulk RNA-seq approaches sample gene expression across millions of cells within a tissue, where allele-specific expression across the X-chromosome can be extracted to model the XCI ratio of the sample. Take a tissue where 75% of cells carry an active maternal X-allele, the other 25% carry an active paternal X-allele. When sampling gene expression in aggregate across this cell population, it is expected approximately 75% of sequencing reads will align to the maternal X-chromosome for any given heterozygous locus. However, there are many factors that can influence the allele-specific expression of any given individual gene, where the allelic expression ratio of a single gene may not reflect the true XCI ratio of the bulk tissue.

Lowly expressed genes will have increased variability at the read-sampling level, identical to the relationship between small cell counts and increased XCI ratio variance (binomial sampling). Expression quantitative trait loci (eQTLs) can also cause deviations in allelic-expression from the underlying XCI ratio. And, escape from inactivation can also affect allele-specific expression of individual genes, a phenomenon where some genes are expressed from the inactive X-allele, producing biallelic expression in cells (Carrel and Willard 2005; Tukiainen et al. 2017). Luckily, XCI occurs on a chromosome-wide scale and aggregating allele-specific expression across numerous well-expressed genes averages in favor of the underlying XCI ratio of the tissue. Special attention is given to escape from inactivation, which may affect between 15-30% of the X-chromosome and also may be individual and/or tissue-specific (Berletch et al. 2015; Tukiainen et al. 2017; Zito et al. 2021). In Aim 1, previously annotated human escape genes are excluded from analysis and in Aim 2, regions of the X-chromosome that exhibit escape signal are excluded across species.

The central approach for modeling XCI ratios from allele-specific gene expression is to first identify heterozygous loci and then determine the number of reads that align to each parental allele. This typically requires the use of additional genetic information in the form of personal and/or parental DNA-sequencing to identify heterozygous variants and determine the phasing of each read (allelic identity of each read). Such approaches are costly and do not enable scalable analysis of XCI ratios, a requirement for cross-species assessments specifically. Instead, the combination of identifying heterozygous variants from RNA-sequencing reads and using a reference genome for alignment and quantifying allele-specific expression would allow XCI ratios to be extracted from any bulk RNA-seq sample. Specific methodological details for this approach are provided in the methods sections pertaining to Aim 1 and Aim 2 (pages 48-50, pages 82-83). Briefly, we utilize folded statistical distributions (Urbakh 1967; Gart 1970) to estimate the magnitude of XCI ratios through aggregated allelic-expression of numerous X-linked genes. This represents a scalable method of assessing XCI ratios applicable to any bulk RNA-seq sample, enabling both large-scale cross-tissue and cross-species analysis of XCI ratios.

1.6. Genetic considerations for variability in XCI ratios

So far, variability in XCI ratios has been introduced through the lens of developmental lineage and stochasticity. However, there exists credible evidence for genetic influences on XCI ratios that may impact XCI ratio variability outside of developmental stochasticity. These can be summarized as either direct genetic effects on XIST that can influence the initial choice of allelic inactivation or mechanisms of allelic-selection through genetic variability that can operate across development and introduce variability into XCI ratios (Migeon 1998; Brown and Robinson 2000).

The primary example of clear genetic influence on XCI ratios is the well-described preferential inactivation of specific X-alleles in heterozygous laboratory mice. Extensive genetic and molecular analysis of XCI in mice have revealed a specific X-linked locus involved in the regulation of XIST expression and subsequently XCI, termed the X-chromosome controlling element (XCE) (Simmler et al. 1993; Calaway et al. 2013; Sun et al. 2021). Heterozygous crosses exhibit preferential allelic X-inactivation dependent on the XCE alleles carried by the parental strains. To date, there is evidence for 6 such XCE-alleles that together form an allelic series of inactivation (Sun et al. 2021). The presence of a genetic locus with direct influence on the choice of allelic-inactivation in mice suggests the same may be true in human populations, due to the fact XCI is a highly conserved feature across mammalian species. However, evidence for a comparable XCE-locus in humans is so far non-existent (Peeters et al. 2016), with the strongest evidence for direct genetic effects on XCI in humans limited to select variants found in/around XIST, typically restricted to small family studies (Plenge et al. 1997). It remains an open question whether XCE-like effects exist in other mammalian species, as the vast majority of XCI data is derived from human and mouse populations.

A separate case for genetic influence over XCI ratios is the active selection for or against specific X-alleles over developmental timespans, instigated through genetic variability across the X-alleles (Migeon 1998; Belmont 1996). Such genetic variability can be disease-related, with a disease variant imparting a selective effect, or through an aggregate effect of chromosome-wide heterozygosity across the alleles, though evidence for the latter is weak. For the disease-case, there is extensive evidence for such selection in humans across distinct X-linked diseases. Genetic disorders that impact the proliferation rates of carrier cells can drive skewing in favor of the mutant allele, manifesting in disease phenotypes for heterozygous carriers (Migeon 1971; Migeon et al. 1981; Devriendt et al. 1997). Similarly, if not always disease associated, large-scale structural aberrations, such as X-autosome translocations, have evidence for instigating allelic selection and result in skewed XCI (Schmidt and Sart 1992). Typically, such cases of allelic-selection manifest through extremely skewed XCI ratios. While this association has linked genetic influence to XCI ratios within the literature, it cannot explain the observed continuous nature of XCI ratios in normal populations.

1.7. Utility of cross-species analysis of XCI ratios

For assessing the relative influences of stochastic or genetic effects on XCI ratios, cross-species analysis is particularly powerful. To date, the vast majority of data on population XCI ratio variability is derived from human and mouse populations, making it difficult to distinguish between species-specific or conserved aspects of XCI. Our approach for modeling XCI ratios from RNA-seq reads exploits naturally occurring genetic variability, enabling broad assessments of genetic associations with XCI ratios, relevant to any species analyzed. In Aim 2, we first establish population distributions of XCI ratios across 9 mammalian species and assess the explanatory power of stochastic models. We then explore potential genetic associations with XCI ratios, at both a broad chromosomal and variant-specific level.

1.8. Summary for Aims 1 and 2

In summary for aims 1 and 2, variability in XCI is highly informative for inferring characteristics of developmental lineage at multiple scales, but has yet to be fully exploited for the study of human development and cross-species assessments. Utilizing an approach for estimating XCI ratios from reference aligned bulk RNA-seq data, extensive cross-tissue assessments of XCI variability are performed using the GTEx dataset and population variability in XCI ratios is computed across 9 mammalian species. This work characterizes early lineage specification events during human development and explores the stochastic and genetic basis of XCI variability across mammals, extending our understanding of XCI and its developmental consequences on multiple fronts.

1.9. Organoid systems for *in vitro* modeling of developmental lineage

Experimental study of *in vivo* developmental processes is challenging due to the technical difficulties in accessing developing tissues and, specifically for humans, the ethical limitations for performing experiments on embryonic material. *In vitro* cell culture approaches have been an instrumental tool in the experimental study of developmental lineage, where the process of differentiation can be captured and interrogated within a dish. An exciting recent methodological development was the advent of 3-dimensional cell culturing systems, where stem cells are cultured within a 3D scaffold of extracellular matrix (Eiraku et al. 2008; Sato et al. 2009). The shift to a 3D culture environment brought remarkable changes to the behavior and lineage production of stem cells, which were able to now produce multi-cellular structures composed of a variety of cell types with spatial organization resembling endogenous tissues (Lancaster et al. 2013; Corrò et al. 2020). These self-organized structures are termed organoids and present the opportunity for experimental *in vitro* study of developmental lineage in a medium much closer to *in vivo* development compared to traditional 2D cell culture. An impressive array of tissue-specific organoids has been established over the years, including lung (Sachs et al. 2019), kidney (Takasato et al. 2015), liver (Huch et al. 2013), intestinal (Sato et al. 2009), and neural organoids (Watanabe et al. 2005; Lancaster et al. 2013) to name a few. This diversity in model systems holds promise to rapidly expand our understanding across an extensive range of developmental phenomenon. However, observations in organoids are only applicable to *in vivo* developmental processes depending on how accurately these models recapitulate primary (*in vivo*) tissue development, an area of active investigation across organoid systems.

The practice of comparing *in vivo* and *in vitro* data to assess similarities and differences (Camp et al. 2015; Velasco et al. 2019; Bhaduri et al. 2020; Gordon et al. 2021; Feng et al. 2022) is technically challenging, as these comparisons are inherently confounded by batch (Leek et al. 2010). This makes it difficult to disentangle batch effects from the underlying primary tissue and organoid biology, raising concerns on the generalizability of findings from such comparisons. While numerous batch integration techniques exist for a variety of batch scenarios (Cheroni et al. 2022), integration can strip away meaningful biological variation in the effort to standardize across batches (Zhang et al. 2023). An independent approach to integration is the meta-analytic assessment of replicability across batches (Tanaka et al. 2020; Cheroni et al. 2022; Kim et al. 2023), identifying the biological signal that is robust to batch effects and likely a more comprehensive representation of the underlying biology. For primary tissue and organoid comparisons, this would entail first identifying replicable signal across batches independently for primary tissue and organoid datasets and then assessing the degree organoids recapitulate primary tissue biology. An informative biological signal for this purpose that also enables the assessment of developmental lineage across primary tissue and organoids is gene co-expression (Zhang and Horvath 2005).

1.10. Gene co-expression for functional comparisons across biological systems

A long-standing observation is that genes which are functionally related are expressed together, where the strength of correlated gene expression defines the co-expression relationship between genes (Stuart et al. 2003). Gene co-expression is particularly amenable to meta-analysis as it defines a shared genomic space that can be aggregated across batches (Lee et al. 2020), identifying the functional genomic output that is robust to batch effects. This is especially relevant to primary tissue and organoid comparisons, where gene co-expression can quantify functional similarities and differences between the systems.

The co-expression relationships of genes are ultimately a product of the epigenetic configuration of the genome, which directs which portions of the genome are active to produce some phenotype. This in turn is a product of the developmental lineage of the genome, the carefully orchestrated genomic reconfigurations that transition from a pluripotent state to some specific phenotypic end-state. Taken together, comparisons of gene co-expression across primary tissue and organoid systems can be interpreted as a measure of comparable developmental lineage, quantifying whether lineages *in vitro* were successful in producing the same endo-phenotypes as primary tissue.

1.11. Neural organoids

A particularly challenging developmental process to study is the development of the brain, by large the most complex tissue in terms of the diversity of cell-types produced and their spatial arrangements. It follows that neural organoids are one of the more diverse organoid systems, with immense variability across protocols attempting to model ever increasingly specific brain regions or cell-types/lineages (Mayhew and Singhania 2023; Lancaster and Knoblich 2014; Sakaguchi et al. 2015; Qian et al. 2016; Xiang et al. 2017; Birey et al. 2017; Xiang et al. 2019; Miura et al. 2020; Eura et al. 2020; Andersen et al. 2020; Huang et al. 2021; Sozzi et al. 2022). This technical diversity across neural organoids complicates assessments of fidelity to primary tissues, where each protocol is assessed *ad hoc* in a study-specific manner with few considerations for generalizability across protocols. Meta-analytic approaches stand to be particularly useful for neural organoids, potentially identifying replicable features that can be exploited for generalizable quality control metrics. A more detailed introduction to neural organoids is provided in section 4.4.

There are many axes of variation to consider when comparing primary neural tissue and neural organoids, which are often approached with attempts to isolate specific biological variability as much as possible to make controlled observations. In other words, the goal of neural organoids is to accurately model the regional, temporal, and genetic dynamics that produce specific cell-types, so regional, temporal, and genetic cell-type controls are typically employed when comparing to primary neural tissues. Rather than ever increasingly specific controls that might be difficult to define or capture in primary neural tissues, more general representations of cell-types across brain regions, timepoints, and individuals would facilitate a more generalizable framework for *in vivo*/*in vitro* comparisons.

1.12. Summary for Aim 3

In the third aim of this thesis, meta-analytic co-expression of specific cell-types is derived across highly diverse primary tissue and organoid datasets, sampling across the first two trimesters of neural development, numerous brain regions and individuals, and sampling numerous organoid differentiation protocols. We argue primary tissue cell-type specific co-expression relationships that are robust to temporal, regional, and technical variation constitute a quality control benchmark applicable to any neural organoid dataset for *in vivo*/*in vitro* assessments of fidelity. Specifically, we focus on quantifying the preservation of co-expression across systems, which provides quantifications of fidelity at the gene, cell-type, and whole genome scales. In summary, this work presents a field-wide assessment of fidelity between *in vitro* models of neural development and primary neural tissues, deriving a generalizable quantitative benchmark applicable to highly heterogeneous neural organoid systems.

2. Variability of cross-tissue X-chromosome inactivation characterizes timing of human embryonic lineage specification events

2.1. Citation

Werner JM, Ballouz S, Hover J, Gillis J. 2022. Variability of cross-tissue X-chromosome inactivation characterizes timing of human embryonic lineage specification events. Developmental Cell 57: 1995-2008.e5.

2.2. Author contributions and Acknowledgements

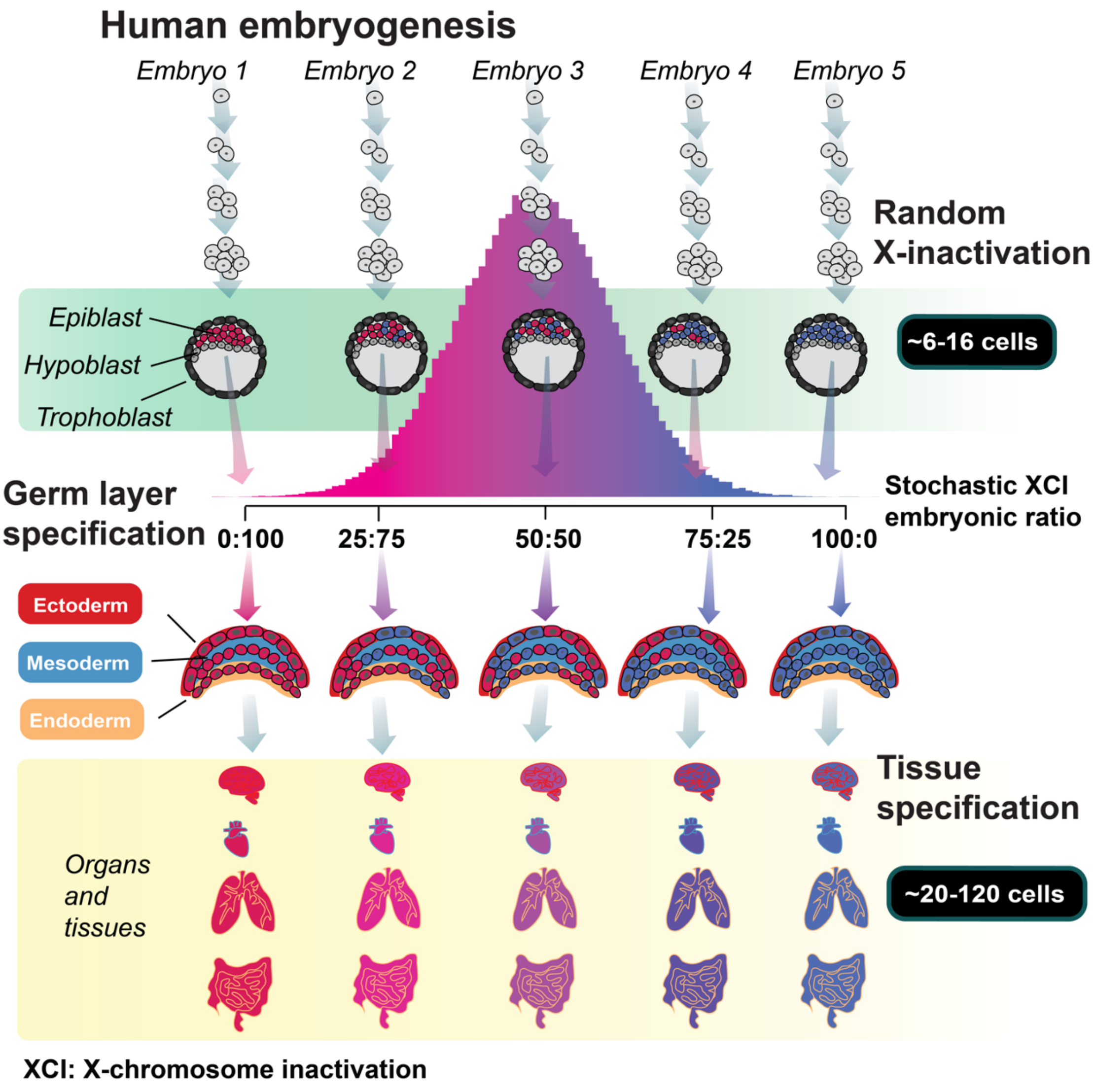
J.G. conceived the project. J.M.W. and J.G. designed the experiments and wrote the manuscript. J.M.W. performed the experiments. S.B. performed preliminary data analysis. S.B. and J.H. assisted with data management and other initial data curation.

JMW was supported by NSF Award No. DGE-1938105. JG, SB and JH were supported by NIH Grants R01MH113005 and R01LM012736. This material is based upon work supported by the National Science Foundation Graduate Research Fellowship Program under Grant No. DGE-1938105. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation. We thank all members of the Gillis lab and particularly Manthan Shah for assisting in some of the initial downloading.

2.3. Results summary

In this work, human cross-tissue variability of XCI ratios is assessed using bulk RNA-seq samples from the GTEx dataset, sampling across 49 different tissues from 311 individuals, representing all 3 germ layer lineages. Quantifying the degree of variance in XCI ratios within lineages or shared variance across lineages is informative for inferring progenitor pool cell counts and lineage relationships. The breadth of tissue sampling within the GTEx dataset presents the opportunity for investigating organism-wide lineage relationships and for resolving characteristics of early developmental lineage in humans, which are otherwise difficult to assess. First, using phased data from the EN-TEx consortium, we demonstrate the accuracy of our approach for estimating XCI ratios from bulk reference-aligned RNA-seq samples. We then explore the consequences of escape from XCI on our modeling results, revealing negligible impacts, and provide novel evidence of escape for 19 genes. We compute correlations of XCI ratios across all tissues, demonstrating that XCI ratios are comparably shared across tissues derived from all three germ-layers. We additionally deconvolve the bulk RNA-seq samples into germ-layer specific contributions using single-cell RNA-seq from the GTEx dataset, revealing XCI ratios are correlated across germ-layer specific markers, corroborating results from the non-deconvolved data. We estimate the number of cells that must have been present during XCI in the embryonic epiblast to be between 6-16 cells to produce the observed variance in XCI ratios across tissues. We additionally explore lineage-specific variability in XCI ratios to model cell counts for tissue-specific lineage specification. We demonstrate a subset of tissues are enriched for switching the dominant parental direction of inactivation. This suggests increased variance in XCI ratios within specific lineages, which we explain through a model of small cell number sampling during lineage specification. In conclusion, this work demonstrates XCI ratios are generally shared across all tissues, providing evidence that XCI occurs before any tissue specification in human development and the stochastic embryonically determined XCI ratio is propagated through development to all tissues. While additional cell sampling events can contribute to XCI variability for specific tissues, we conclude much of the observed variation in XCI ratios within human populations can be explained by the inherent stochasticity of XCI. This work resolves early organism-wide lineage relationships and infers lineage-specific characteristics during human development by exploiting the early, random, and permanent nature of XCI.

2.4 Graphical abstract:



2.5 Introduction:

Every cell within female mammalian embryos undergoes the process of X-chromosome inactivation (XCI), which silences expression from a single randomly chosen X-allele via epigenetic mechanisms (Migeon 2013; Lyon 1961; Dossin and Heard 2021). The random choice of which allele to inactivate occurs early in development and is permanent thereafter with the inactivated allele propagated through each cell's developmental lineage(Lyon 1972). As a result, adult females exhibit mosaic X-linked allelic expression throughout every tissue within the body, an enduring phenotypic consequence of an early embryonic milestone. The random, permanent, and developmentally early nature of XCI positions the whole-body mosaicism of X-linked allelic expression as a lineage marker reaching back to the earliest embryonic stages (Mclaren 1972; Nesbitt 1971). Careful analysis of X-linked allelic expression across individuals and tissues can thus reveal whole-body lineage relationships stemming from some of the first lineage decisions made during embryogenesis (Nesbitt 1971; Fialkow 1973; Bittel et al. 2008; Monteiro et al. 1998).

While the probability for inactivation is equal between the X-alleles in humans, variation in XCI allelic ratios across individuals is a salient feature of XCI. Deviation from the expected XCI allelic ratio of 0.5 can arise through various mechanisms (Brown and Robinson 2000; Schmidt and Sart 1992; Naumova et al. 1996; Wu et al. 2014) with the most basic being the inherent stochasticity of the initial choice of allelic inactivation (Shvetsova et al. 2019). The variability of the initial XCI ratio within the embryo is directly linked to the number of cells present during inactivation where smaller cell numbers result in increased variability of XCI ratios (Nesbitt 1971). In fact, one can estimate the number of cells present at the time of inactivation by analyzing the variance of XCI ratios across a population. Several studies using this approach (Shvetsova et al. 2019; Amos-Landgraf et al. 2006) , as well as studies utilizing *in vitro* embryonic models (Moreira de Mello et al. 2017; Petropoulos et al. 2016; van den Berg et al. 2009), have estimated that XCI occurs in a small stem cell pool within the human embryo with estimates as little as 8 cells. The combination of the random nature and small pool of cells present during XCI imparts an ever-present basal-level of variability in XCI ratios within adult human populations.

The stability of XCI down lineages means that minor cell sampling variation can be used as a marker for any process involving selection of a set of cells, i.e., lineage specification (Nesbitt 1971; Fialkow 1973). While growing evidence indicates XCI is initiated early (Moreira de Mello et al. 2017; Petropoulos et al. 2016; van den Berg et al. 2009) , the exact timing of XCI as it relates to early lineage specification is unclear (Geens and Chuva De Sousa Lopes 2017) and has important implications for the variance in XCI ratios across early lineages. Specifically, the extent of variability in XCI across adult tissues, those derived from the embryonic lineage during embryogenesis, is a long-standing question (Bittel et al. 2008; Hoon et al. 2015) and directly linked to the timing of XCI and early lineage events. Germ layer specification is the first lineage decision made for all future embryonic tissues and occurs during post-implantation embryonic development (Ghimire et al. 2021), a similar timeframe to XCI. If XCI is completed before germ layer specification each germ layer would be specified from the same pool of cells with a set XCI ratio (Fig. 2.1A). The germ layer-specific XCI ratio would be dependent on the initial XCI ratio resulting in shared XCI ratios across germ layers (Fig. 2.1A) and the subsequently derived adult tissues. In contrast, if XCI is completed after germ layer specification, germ layer-specific XCI ratios are set independently and are not expected to be shared across the different germ layers (Fig. 2.1B), producing variance in XCI ratios across adult tissues. Consequently, comparing XCI ratios for tissues within either the same or different germ layer lineages can reveal the temporal ordering of XCI and germ layer specification.

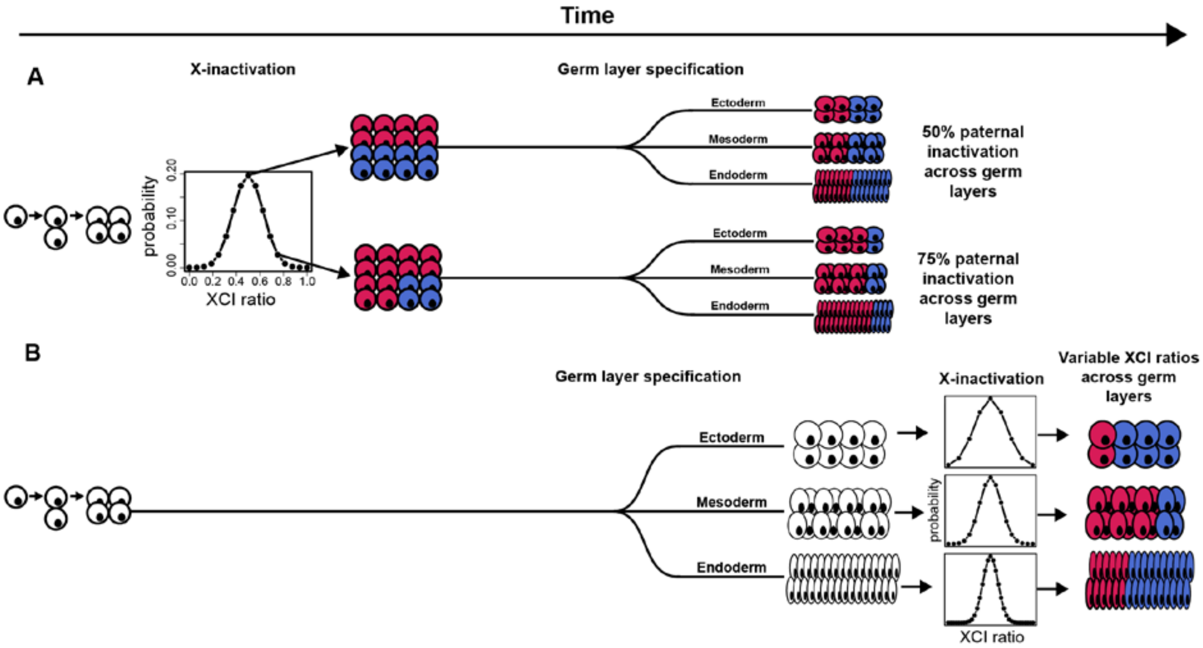


Figure 2.1: ***Timing of XCI determines lineage-specific XCI ratio probability***

A, Schematic representing completed XCI before germ layer specification. Each germ layer inherits the same randomly determined XCI ratio set prior to germ layer lineage specification. The probability distribution of XCI is determined by the number of cells present during inactivation. **B**, Schematic representing completed XCI after germ layer specification. The XCI ratio for each germ layer is set independent of one another, together along with variation in cell numbers fated for each germ layer results in variable XCI ratios across the germ layer lineages.

An additional early lineage event that may overlap with XCI is extraembryonic/embryonic lineage specification (Moreira de Mello et al. 2017; Petropoulos et al. 2016), which precedes germ layer lineage specification. If XCI occurs before or during extraembryonic/embryonic lineage specification, variance in XCI ratios across adult tissues will be influenced by the initial stochasticity of XCI and the subsequent cell selection for the embryonic lineage. In other words, variance in XCI ratios across the germ layer lineages is tied to their last developmental common denominator: the specification of the embryonic epiblast. Since extraembryonic tissues do not contribute to adult tissues, the timing of XCI and extraembryonic/embryonic lineage specification provides the developmental context that variance in adult tissues is potentially tied to the specification of the embryonic epiblast.

In this study, we develop an approach to determine the tissue XCI ratio from unphased bulk RNA-sequencing data, allowing us to assess XCI ratios from any publicly available RNA-sequencing dataset. Utilizing the tissue sampling scheme of the Genotype-Tissue Expression (GTEx v8) project (Lonsdale et al. 2013), we analyze XCI ratios for 49 tissues both within and across individuals for 311 female donors (Fig. S2.1). We establish that XCI ratios are shared for tissues both within and across germ layers demonstrating that XCI is completed before any significant lineage decisions are made for embryonic tissues. Additionally, we extend population-level modeling of variance in XCI ratios to all well-powered tissues, deriving estimates for the number of cells present at the time of embryonic epiblast and tissue-specific lineage commitment. By providing cell counts, temporal ordering of lineage events, and lineage relationships across tissues, capturing the statistical commonalities that underlie the inherently stochastic nature of XCI is a powerful approach for resolving questions of early developmental lineage specification.

2.6 Results

2.6.1 The folded-normal model accurately estimates XCI ratios from unphased data

A practical consequence of bulk RNA-sequencing is that the XCI ratio of a tissue can be estimated from the direction and magnitude of X-linked allele-specific expression. For a tissue with 75% of cells carrying an active maternal X-allele, approximately 75% of RNA-sequencing reads for heterozygous loci are expected to align to the maternal X-allele (Fig. 2.2A). However, allelic expression for any given gene is affected by a variety of factors both biological (e.g., eQTLs) and technical (e.g., read sampling). To derive robust estimates, we aggregate allelic expression ratios across well-powered intra-genic heterozygous SNPs for a given tissue, providing a chromosome-wide estimate of the tissue XCI ratio (Fig, 2.2A).

When aligned to a reference genome, reference alleles will be composed of both maternal and paternal alleles for a given sample. It follows that reference allelic expression ratios represent the expected expression ratios from both the maternal and paternal alleles given the XCI ratio of the tissue (Fig. 2.2A). To account for this, folding the reference allelic expression ratios about 0.5 aggregates the imbalanced allelic expression within the tissue across the two alleles. This enables the magnitude of the XCI ratio to be estimated from unphased expression data by fitting a folded distribution (Gart 1970; Urbakh 1967) (see methods, Fig. 2.2A-B).

To assess the accuracy of the folded-normal model in estimating XCI ratios, we test our approach with phased bulk RNA-sequencing data from the EN-TEx (Rozowsky et al. 2021) consortium, a total of 49 tissue samples from 2 female donors spanning 26 different tissues. Comparing the unphased estimates derived with the folded-normal model to the phased median allelic expression per sample, we find nearly perfect XCI ratio estimate correspondence for ratios greater than 0.6 (Fig. 2.2C). For samples skewed closer to the folding point of 0.5, model misspecification of the underlying distribution makes the estimate overconservative.

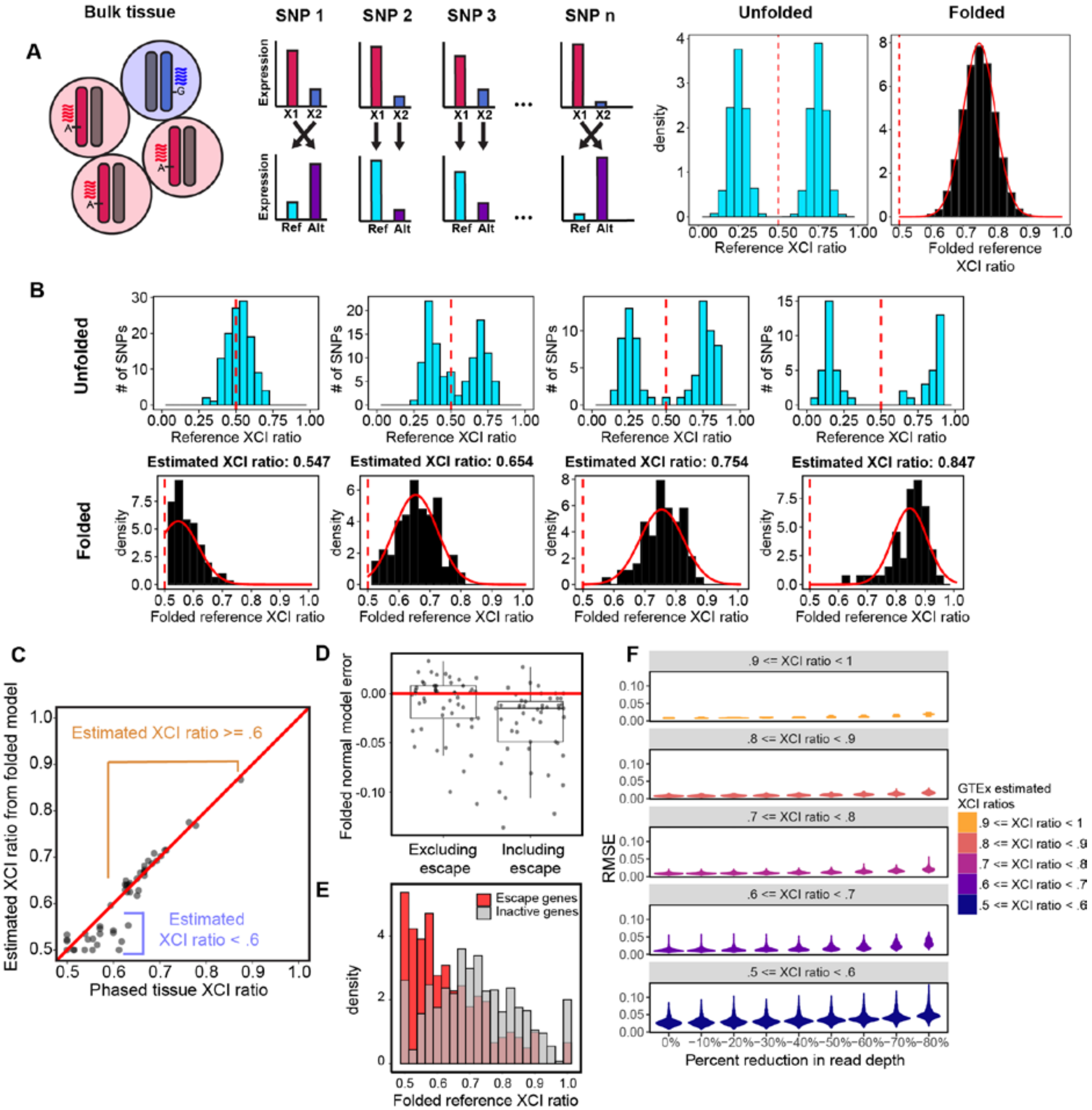

A
Bulk tissue
A
A
A
G
SNP 1
SNP 2
SNP 3
SNP n
Expression
X1 X2
Ref Alt
...
Unfolded
Folded
density
Reference XCI ratio
Folded reference XCI ratio
B
Unfolded
# of SNPs
Reference XCI ratio
Estimated XCI ratio: 0.547
Estimated XCI ratio: 0.654
Estimated XCI ratio: 0.754
Estimated XCI ratio: 0.847
Folded
density
Folded reference XCI ratio
C
Estimated XCI ratio from folded model
Phased tissue XCI ratio
Estimated XCI ratio >= .6
Estimated XCI ratio < .6
D
Folded normal model error
Excluding escape
Including escape
E
Escape genes
Inactive genes
density
Folded reference XCI ratio
F
.9 <= XCI ratio < 1
.8 <= XCI ratio < .9
.7 <= XCI ratio < .8
.6 <= XCI ratio < .7
.5 <= XCI ratio < .6
RMSE
0% -10% -20% -30% -40% -50% -60% -70% -80%
Percent reduction in read depth
GTEx estimated XCI ratios
.9 <= XCI ratio < 1
.8 <= XCI ratio < .9
.7 <= XCI ratio < .8
.6 <= XCI ratio < .7
.5 <= XCI ratio < .6

Figure 2.2: ***The folded-normal model accurately estimates XCI ratios from unphased bulk RNA-sequencing data***

A, Schematic demonstrating how allelic expression of heterozygous SNPs reflect the XCI ratio of bulk tissue samples. Aligning expression data to a reference genome scrambles the parental haplotypes. Folding the reference allelic expression ratios captures the magnitude of the tissue XCI ratio. **B**, Distributions of reference allelic expression ratios for identified heterozygous SNPs across tissue samples exhibiting a range of bulk XCI ratios. Both the unfolded (top row) and folded distributions with the fitted folded normal model (bottom row) are shown. **C**, For the EN-TEx tissue samples, the phased median gene XCI ratio is plotted against the unphased XCI ratio estimate from the folded normal model. The folded normal model produces near identical XCI ratio estimates for samples with XCI ratios greater than or equal to 0.60. **D**, Deviation of the folded normal model from the phased median gene XCI ratio when excluding or including known escape genes. **E**, Aggregated folded reference allelic expression distributions for known escape and inactive genes in EN-TEx tissues with XCI ratios >= 0.70. **F**, Root mean squared error distributions for GTEx tissue samples binned by their original estimated XCI ratio as read depth per SNP is gradually reduced. See also Figure S2.1.

Our approach for estimating XCI ratios aggregates allelic expression across numerous heterozygous loci, averaging away mechanisms outside of XCI that may impact X-linked allelic expression. A widespread mechanism that may still impact our XCI ratio estimates is escape from inactivation, where a gene is biallelically expressed from the active and inactive X-alleles (Tukiainen et al. 2017). Between 15-30% of genes on the X-chromosome have documented evidence for escape (Tukiainen et al. 2017; Carrel and Willard 2005). While we exclude known escape genes (Tukiainen et al. 2017) from our folded-normal XCI ratio estimates, it is very likely unannotated escape genes are present within the data. To identify the impact of escape on our XCI ratio estimates, we compare folded-normal XCI ratio estimates derived with either excluding or including known escape genes to the phased XCI ratio of tissues excluding the known escape genes (Fig. 2.2D). Including known escape genes biases the folded-normal XCI ratio estimates towards 0.5 (Fig. 2.2D). By comparing allelic ratios of known escape genes to all other genes in EN-TEx tissues with XCI ratios >= 0.7, we clearly see escape genes trend towards balanced biallelic expression contributing to the underestimated XCI ratios when including escape genes (Fig. 2.2E).

To assess variance in XCI and escape more broadly, we capitalize on the tissue sampling structure of the Genotype-Tissue Expression (GTEx v8) dataset (Fig. S2.1). From an average of 56 +- 23.5 (SD) well-powered heterozygous SNPs (genes, see methods) per sample (Fig. S2.1), we derive robust XCI ratio estimates for 4658 GTEx tissue samples spanning 49 different tissues (Fig. S2.1).

In addition to biological sources of variation (escape), read depth is a critical source of technical variation to assess when analyzing allelic expression. Sampled allelic expression is the result of a binomial sampling event dependent on the number of reads sampled and the probability of allelic expression. While we employ stringent read count requirements (see methods), we additionally explore how robust our tissue-level XCI ratio estimates are in the face of global decreases in read depths across genes (Fig. 2.2F). As read depths per gene are decreased (10%, 20%, 30%, etc.), the vast majority of increased error in the XCI ratio estimates is constrained to the estimates below 0.6 (Fig. 2.2F), whereas the most skewed tissue samples (XCI ratio estimates above 0.9) display nearly zero additional error even up to an 80% reduction in read depth (Fig. 2.2F). These results are in line with our phased vs unphased comparisons demonstrating XCI ratio estimates above 0.6 (Fig. 2.2C) are highly accurate. Additionally, these results appear to be independent of the number of genes used to estimate the tissue XCI ratio (Fig. S2.1), where we use a minimum of 10 genes per sample. This suggests that aggregating allelic expression over even a modest number of genes is powered to accurately estimate tissue XCI ratios above 0.6 from bulk RNA-sequencing data.

2.6.2 Escape genes exhibit consistent cross-tissue biallelic expression

Our method to quantitatively determine the tissue XCI ratio via aggregating signal across genes is especially well-suited to explore escape from XCI within the GTEx dataset (Fig. 2.2E). Our basic strategy for detecting escape genes is to calculate each gene’s consistency with the aggregate chromosomal inactivation ratio. Assessing all X-linked genes utilized in our GTEx XCI ratio estimates (Fig. 2.3A) and previously annotated constitutively escape genes (Tukiainen et al. 2017) results in a wide range of correlations between gene and tissue XCI ratios, exemplified by the genes SHROOM4 and TCEAL3 (Fig. 2.3B). As expected, the transcripts associated with XCI, namely, XIST and TSIX, show some of the highest correlations to the tissue XCI ratio (i.e., top 8.7%, Fig. 2.3B). Similarly, known escape genes exhibit some of the smallest correlations (Fig. 2.3B). Interestingly, several genes previously annotated as escape do exhibit rather strong correlations to the XCI ratio of tissues. We find that increased gene expression is linked to increased correlation to the tissue XCI ratio (Fig. 2.3C) suggesting that some gene variation with respect to the tissue XCI ratio is technical, reflecting read sampling at low expression. At matched expression levels, previously annotated escape genes have smaller tissue-gene XCI ratio correlations compared to all other genes (Fig. 2.3C), demonstrating that known escape genes are less correlated to the tissue XCI ratio as expected by expression levels alone.

From our analysis in the EN-TEx dataset, escape from inactivation trends toward balanced biallelic expression rather than achieving completely equal allelic expression (Fig. 2.2E), explaining how some escape genes retain significant correlations to tissue XCI ratios in the GTEx dataset. To comprehensively test the degree to which escape produces balanced allelic expression, we construct a one-sided test to detect whether a gene consistently trends towards balanced biallelic expression regardless of the XCI ratio of the tissue (see methods, Fig. S2.2). Against a null distribution of inactivated genes, we are able to identify genes with consistent biallelic expression in opposition to the aggregate imbalanced tissue XCI ratio, indicating escape from XCI (Fig. 2.3D).

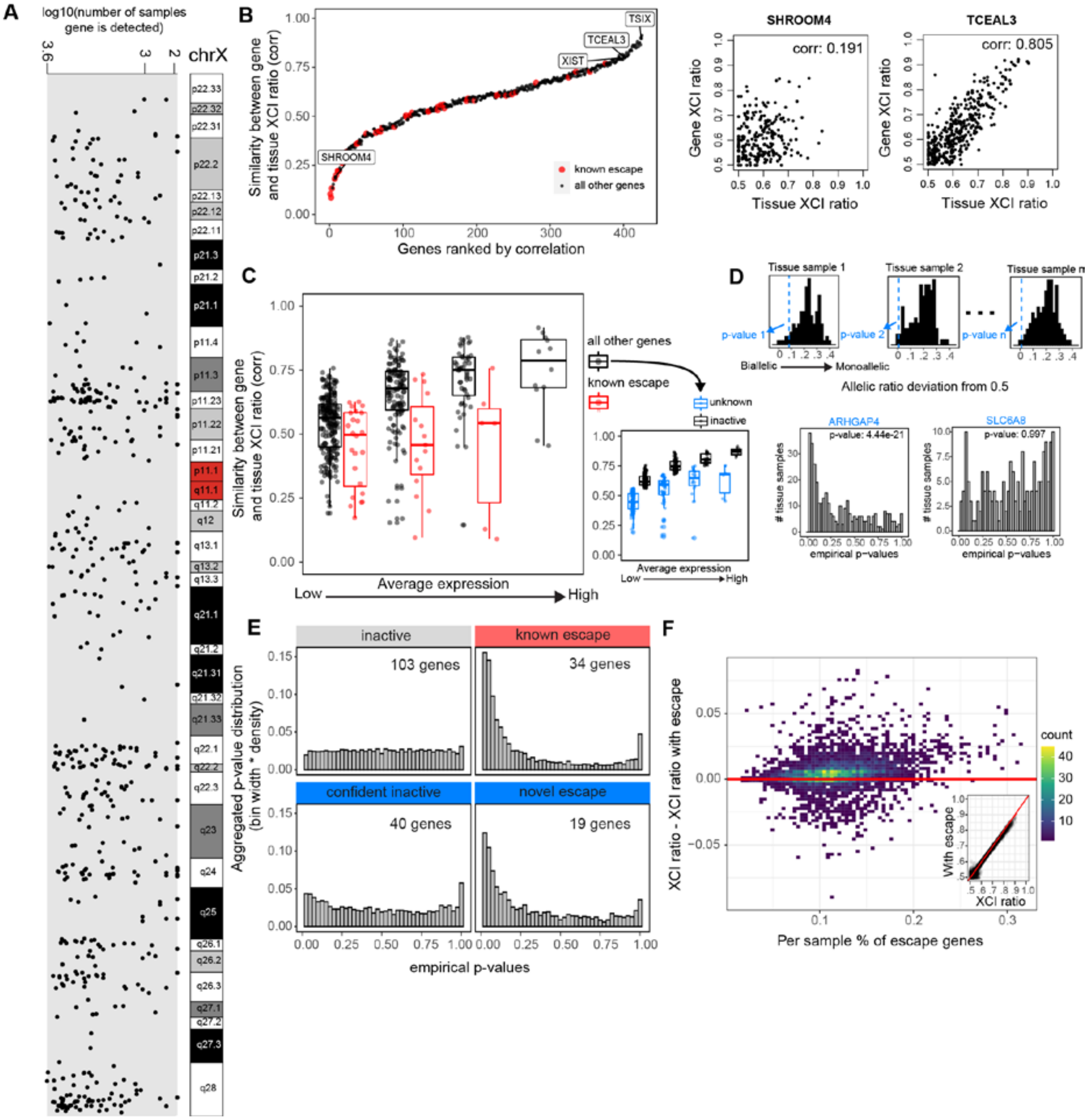

A
log10(number of samples gene is detected)
3.6
3
2
chrX
p22.33
p22.32
p22.31
p22.2
p22.13
p22.12
p22.11
p21.3
p21.2
p21.1
p11.4
p11.3
p11.23
p11.22
p11.21
p11.1
q11.1
q11.2
q12
q13.1
q13.2
q13.3
q21.1
q21.2
q21.31
q21.32
q21.33
q22.1
q22.2
q22.3
q23
q24
q25
q26.1
q26.2
q26.3
q27.1
q27.2
q27.3
q28
B
Similarity between gene and tissue XCI ratio (corr)
Genes ranked by correlation
TSIX
TCEAL3
XIST
SHROOM4
known escape
all other genes
SHROOM4
corr: 0.191
TCEAL3
corr: 0.805
Gene XCI ratio
Tissue XCI ratio
C
Similarity between gene and tissue XCI ratio (corr)
Average expression
Low
High
all other genes
known escape
unknown
inactive
Average expression
Low
High
D
Tissue sample 1
Tissue sample 2
Tissue sample n
p-value 1
p-value 2
p-value n
Biallelic
Monoallelic
Allelic ratio deviation from 0.5
ARHGAP4
p-value: 4.44e-21
SLC6A8
p-value: 0.997
# tissue samples
empirical p-values
E
Aggregated p-value distribution (bin width * density)
inactive
103 genes
known escape
34 genes
confident inactive
40 genes
novel escape
19 genes
empirical p-values
F
XCI ratio - XCI ratio with escape
Per sample % of escape genes
count
With escape
XCI ratio

Figure 2.3: ***Genes that escape XCI exhibit balanced biallelic expression across XCI skewed tissues***
A, The genomic location and number of GTEx samples each gene is detected for the 542 genes that pass our quality control filters. **B**, All 542 genes and 45 known escape genes ranked by the Pearson correlation coefficient for each gene's allelic expression and the XCI ratio of the tissue for samples that detect that gene. **C**, Distributions of gene-tissue XCI ratio correlations for all 542 genes and 45 escape genes, binned by average expression. The range of average expression is binned into 4 equally spaced bins. We label the top 50% of 'all other genes' in each expression bin as 'inactive genes' and the bottom 50% as 'unknown' genes, as they are potentially a mix of inactive and unannotated escape genes. **D**, An example for how the empirical p-values are calculated for a given test gene across tissue samples. For a given tissue sample, we calculate each gene's allelic expression ratio deviation from 0.5, where the black histogram represents the deviations from the inactive genes in the sample and the blue dotted line represents the deviation of the given test gene in the sample, ARHGAP4 in this example. We apply Fisher's method to aggregate each test gene's distribution of empirical p-values to calculate a meta-analytic p-value to determine significance (ARHGAP4 meta-analytic p-value: 4.44e^{-21}, SLC6A8 meta analytic p-value: 0.997). **E**, The aggregated empirical p-value distributions for inactive, known escape, and the unknown genes now classified as confident inactive and novel escape are plotted. The unknown genes are classified as either confident inactive or novel escape by using a significance threshold of meta-analytic p-value < .001. **F**, The percent of genes previously annotated for escape per sample is plotted against the difference between the sample's XCI ratio estimates derived when either including or excluding the previously annotated escape genes. The inset plot compares the XCI ratio estimates derived without the known escape genes (x-axis) or including the known escape genes (y-axis). See also Figure S2.2 and Table S2.1.

Testing the known escape genes using this approach results in significant escape signal (Fig. 2.3E). Similarly, we are able to identify 19 genes previously unannotated for constitutive escape to have significant escape signal (p-value < .001): ARHGAP4, BTK, CASK, CHRDL1, CLIC2, COX7B, CTPS2, CXorf36, F8, ITM2A, MECP2, MPP1, NLGN4X, PGK1, RPL36A, SASH3, SEPT6, STARD8, VSIG4 (Fig. 2.3E, Fig. S2.4). Revisiting these genes within the literature, several have prior evidence for escape, though typically limited in the tissues assessed: BTK (Hagen et al. 2020; Zito et al. 2021), CASK (Zito et al. 2021), CHRDL1 (Zito et al. 2021), CLIC2 (Tukiainen et al. 2017; Zito et al. 2021), COX7B (Larsson et al. 2019), CTPS2 (Balaton et al. 2021), CXorf36 (Winham et al. 2019), MPP1 (Zito et al. 2021), NLGN4X (Tukiainen et al. 2017; Zito et al. 2021), SASH3 (Zito et al. 2021), SEPT6 (Zhang et al. 2013), VSIG4 (Berletch et al. 2015). Our results suggest these genes escape inactivation more broadly than previously reported. In addition, our analysis provides supporting evidence of escape for 34 previously annotated escape genes and supporting evidence of inactivation for 143 genes (Table S2.1). While in this analysis we are powered to identify more constitutively escape genes, variability in escape across tissues and individuals is well documented. As such, our escape annotations are robust to the GTEx data we sample over and will benefit greatly from future experimental follow up.

To test the impact of including escape genes on our GTEx tissue XCI ratio estimates, we compare our original tissue XCI ratio estimates to estimates calculated while including the known escape genes (Fig. 2.3F). The inclusion of escape genes results in slightly underestimated XCI ratios (Fig. 2.3F), though the impact is minimal with an average absolute deviation of 0.0088 (± 0.010 SD) between XCI ratio estimates including/excluding the known escape genes. This demonstrates our folded aggregation of allelic expression across genes to estimate XCI ratios is robust to noise generated by escape from inactivation.

2.6.3 XCI is completed prior to germ layer specification

Having developed a robust approach to measure XCI ratios from unphased data, we turn to assessing the degree XCI ratios are shared across tissues within individuals. As an initial visualization of XCI ratios across tissues, we order all female GTEx donors by their average XCI ratio and plot the ratio for all tissues grouped by germ layer (Fig. 2.4A). XCI ratios qualitatively appear consistent across all tissues and the three germ layers (Fig. 2.4A). We then ask how well do individual tissues predict all other tissues' XCI ratios, which we quantify with the AUROC (area under receiver operating characteristic curve) metric (Fig. S2.3). For a given tissue, we take the average XCI ratio of all other tissues for each donor and use this average to classify the donors as low/high XCI ratio donors. If the given tissue's XCI ratio can recapitulate the same low/high classifications of the donors, this indicates that tissue's XCI ratio is in concordance with the average of all other tissues and would result in an AUROC close to 1. Across various thresholds for defining low/high donors, we see that performance is high and consistent across all tissues, suggesting XCI ratios are generally shared across all tissues for an individual (Fig. S2.3).

Stratifying tissue comparisons of XCI ratios by germ layer lineage relationships should resolve the temporal ordering of XCI and germ layer specification within the human embryo. If XCI occurs before germ layer specification, tissue XCI ratios are expected to positively covary across tissues from different germ layer lineages (Fig. 2.1A). In contrast, if XCI occurs after germ layer specification, the XCI ratio of each germ layer is set independently and there is little expected covariance in XCI ratios for tissues from different germ layers (Fig. 2.1B). We compute correlations of the XCI ratio for combinations of tissues derived from either the same or different germ layers, exemplified in Figure 2.4 panel B. Tissues sharing the same germ layer lineage produce strictly positive significant correlation values ranging from 0.25 to 0.90 (Fig. 2.4C), demonstrating XCI ratios are shared within individual germ layer lineages. Strikingly, significant positive ratio correlations for tissues derived from different germ layers are on the same order as the within germ layer comparisons, ranging from 0.24 to 0.87 (Fig. 2.4C, Fig. S2.3). The fact tissues derived from different germ layers covary for their XCI ratio strongly suggests XCI is completed prior to germ layer specification and the initial embryonic XCI ratio is propagated through all germ layer lineages.

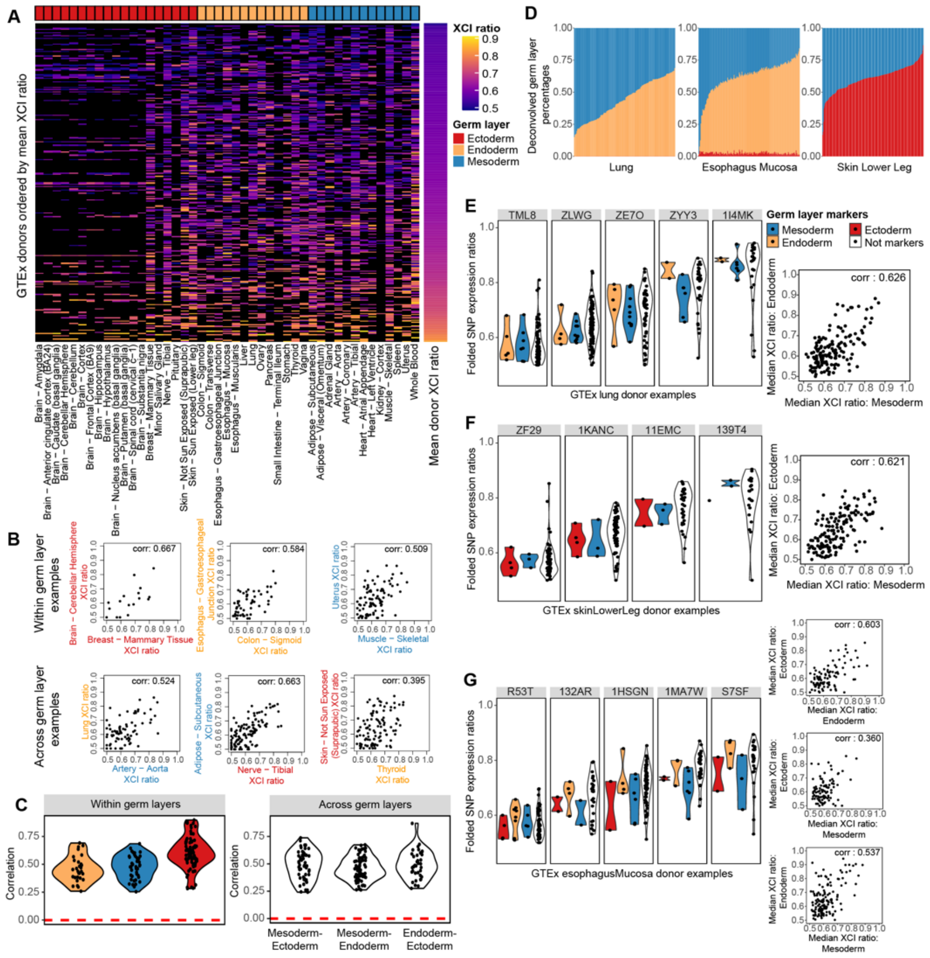


Figure 2.4: ***XCI ratios are shared across germ layer lineages***
A, Heatmap of all estimated XCI ratios for the tissues of each donor, with donors ordered by their mean XCI ratio across tissues and tissues grouped by germ layer lineage. Black indicates no tissue donation for that donor-tissue pair. **B**, Examples of within and across germ layer lineage comparisons of XCI ratios. Each data point represents the estimated XCI ratios of the two indicated tissues for a single donor. **C**, All significant (FDR corrected p-value <= 0.05, permutation test n = 10000) Pearson correlation coefficients for within and across germ layer lineage comparisons. **D**, Stacked bar plots for the germ layer percentage composition for each sample in the Lung, Esophagus Mucosa, and Skin Lower Leg GTEx tissues. The deconvolved cell type percentages and their germ layer annotations are provided in Fig. S2.2. **E-G**, the folded allelic expression ratios for germ layer markers and all other genes (Not markers) are plotted for several example donors per tissue, E: Lung, F: Skin Lower Leg, G: Esophagus Mucosa. The adjacent scatter plots compare the median folded allelic expression between germ layer markers for all donors. E: Lung mesodermal and endodermal markers, Pearson correlation of 0.626 (p-value < .001), F: Skin Lower Leg mesodermal and ectodermal markers, Pearson correlation of 0.621 (p-value < .001), G: Esophagus Mucosa endodermal and ectodermal markers, Pearson correlation 0.603 (p-value < .001), mesodermal and ectodermal markers, Pearson correlation 0.360, (p-value < .001), mesodermal and endodermal markers Pearson correlation 0.537 (p-value < .001). See also Figure S2.3-2.4 and Table S2.2.

While we annotate individual tissues to belong to a single primary germ layer, tissues are compositions of cell types derived from different germ layers. This may impact the observed variance in XCI ratios across tissues if there is a strong germ layer-specific effect in XCI ratio variance. We take advantage of the recently released single-nucleus RNA-sequencing (Eraslan et al. 2022) GTEx data to deconvolve (Newman et al. 2019) several of the bulk tissues into their germ layer components, allowing us to explore variance in XCI ratios across germ layers within single tissues. Figure 2.4D provides examples of the deconvolved germ layer proportions of three tissues with the remaining 6 tissues provided in Figure S2.4, demonstrating there is variation in germ layer composition within tissues. We extract germ layer-specific markers for the lung, skin, and esophagus mucosa tissues (Table S2.2, see methods) to explore variance in XCI ratios across germ layers within single tissues. The XCI ratios of germ layer-specific markers positively covary in each tissue (Fig. 2.4E-G, Pearson correlations: lung mesoderm and endoderm 0.626, skin mesoderm and ectoderm 0.621, esophagus endoderm and ectoderm 0.603, esophagus mesoderm and ectoderm 0.360, esophagus mesoderm and endoderm 0.537), recapitulating the result of shared XCI ratios across germ layers we demonstrate with the non-deconvolved tissues.

2.6.4 Specific tissue lineages have increased probability for switching the parental direction of XCI

In addition to demonstrating that XCI ratios are broadly shared across all tissues, our cross-tissue analysis reveals there is a degree of variability in XCI ratios across tissues within individuals. Comparing distributions of gene-level allelic expression across tissues for individual donors reveals there are often individual tissues that exhibit divergence in XCI ratios in opposition to the general trend of shared XCI ratios (Fig. 2.5A-B). This is evidenced by the divergent distributions of gene-level allelic-expression for the Whole Blood, Vagina, and Skin tissues in donor 11P81 (Fig. 2.5A), and the Esophagus – Mucosa, Vagina, and Skin tissues in donor 1J1OQ (Fig. 2.5B). The presence of individual tissues exhibiting divergent XCI ratios within an individual suggests there may be lineage-specific effects contributing to variance in XCI ratios across tissues.

To further investigate the degree of variation in XCI ratios across tissues, we take advantage of the cross-tissue sampling of individual donors to determine the parental direction of XCI. If an expressed heterozygous SNP is captured for two different tissues of an individual, the reference allele is on the same haplotype and maintains directional allelic information. Thus, calculating the correlation of reference SNP allelic ratios for shared SNPs between two tissues can reveal whether those tissues share the same XCI direction (Fig. 2.5C-D, see methods). When examining a donor with generally high XCI ratios across all tissues (Fig. 2.5C Donor 11P81), we find that all tissues share the same parental direction in allelic inactivation. Whereas a less skewed donor (Fig. 2.5D Donor 1J1OQ, Ovary and Vagina tissues) exhibits a subset of tissues with opposite parental inactivation compared to the majority of tissues for that donor. Across all donors, as the average XCI ratio of their tissues increases, the proportion of their tissues exhibiting switched parental XCI decreases (Fig. 2.5E), with the most skewed donors exhibiting zero tissues with switched parental XCI (Fig. 2.5E). Interestingly, switching parental direction of XCI is in fact concentrated in a subset of tissues, with 12 out of 49 tissues being significantly enriched for instances of switched XCI (Fig. 2.5F, fisher’s exact test, p-value <= 0.5). The existence of individual tissues with increased probability for switching parental directions of XCI is indicative of increased variance in XCI ratios for those particular tissue lineages. We explore this model further in the Results section ‘Cell population estimates at the time of tissue-specific lineage commitment’.

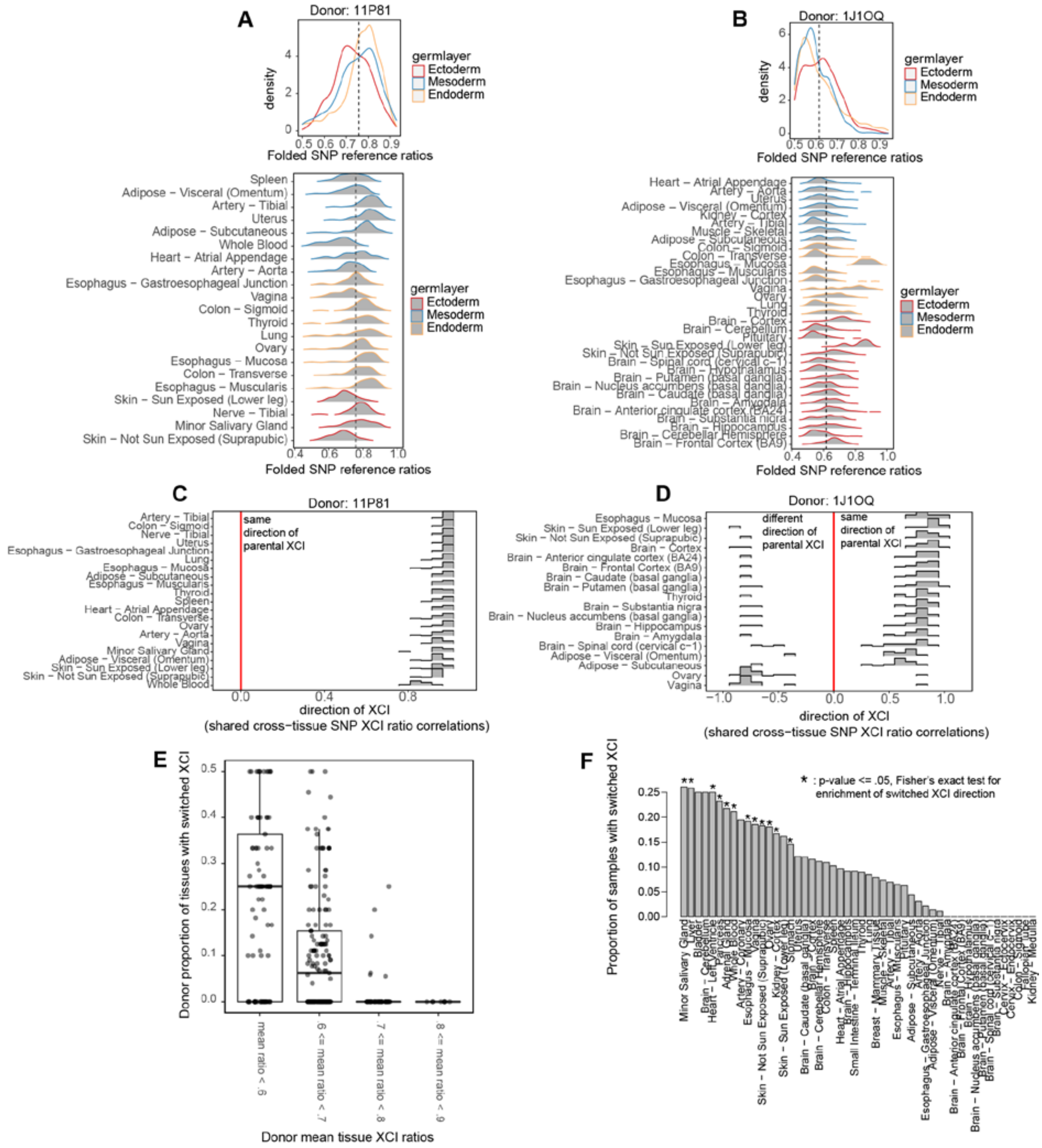

A
Donor: 11P81
density
4
2
0
0.5 0.6 0.7 0.8 0.9
Folded SNP reference ratios
germlayer
Ectoderm
Mesoderm
Endoderm
Spleen
Adipose − Visceral (Omentum)
Artery − Tibial
Uterus
Adipose − Subcutaneous
Whole Blood
Heart − Atrial Appendage
Artery − Aorta
Esophagus − Gastroesophageal Junction
Vagina
Colon − Sigmoid
Thyroid
Lung
Ovary
Esophagus − Mucosa
Colon − Transverse
Esophagus − Muscularis
Skin − Sun Exposed (Lower leg)
Nerve − Tibial
Minor Salivary Gland
Skin − Not Sun Exposed (Suprapubic)
0.4 0.6 0.8 1.0
Folded SNP reference ratios
germlayer
Ectoderm
Mesoderm
Endoderm
B
Donor: 1J1OQ
density
6
4
2
0
0.5 0.6 0.7 0.8 0.9
Folded SNP reference ratios
germlayer
Ectoderm
Mesoderm
Endoderm
Heart − Atrial Appendage
Artery − Aorta
Uterus
Adipose − Visceral (Omentum)
Kidney − Cortex
Artery − Tibial
Muscle − Skeletal
Adipose − Subcutaneous
Colon − Sigmoid
Colon − Transverse
Esophagus − Mucosa
Esophagus − Muscularis
Esophagus − Gastroesophageal Junction
Vagina
Ovary
Lung
Thyroid
Brain − Cortex
Brain − Cerebellum
Pituitary
Skin − Sun Exposed (Lower leg)
Skin − Not Sun Exposed (Suprapubic)
Brain − Spinal cord (cervical c−1)
Brain − Hypothalamus
Brain − Putamen (basal ganglia)
Brain − Nucleus accumbens (basal ganglia)
Brain − Caudate (basal ganglia)
Brain − Amygdala
Brain − Anterior cingulate cortex (BA24)
Brain − Substantia nigra
Brain − Hippocampus
Brain − Cerebellar Hemisphere
Brain − Frontal Cortex (BA9)
0.4 0.6 0.8 1.0
Folded SNP reference ratios
germlayer
Ectoderm
Mesoderm
Endoderm
C
Donor: 11P81
same direction of parental XCI
Artery − Tibial
Colon − Sigmoid
Nerve − Tibial
Uterus
Esophagus − Gastroesophageal Junction
Lung
Esophagus − Mucosa
Adipose − Subcutaneous
Esophagus − Muscularis
Thyroid
Spleen
Heart − Atrial Appendage
Colon − Transverse
Ovary
Artery − Aorta
Vagina
Minor Salivary Gland
Adipose − Visceral (Omentum)
Skin − Sun Exposed (Lower leg)
Skin − Not Sun Exposed (Suprapubic)
Whole Blood
0.0 0.4 0.8
direction of XCI
(shared cross−tissue SNP XCI ratio correlations)
D
Donor: 1J1OQ
different direction of parental XCI
same direction of parental XCI
Esophagus − Mucosa
Skin − Sun Exposed (Lower leg)
Skin − Not Sun Exposed (Suprapubic)
Brain − Cortex
Brain − Anterior cingulate cortex (BA24)
Brain − Frontal Cortex (BA9)
Brain − Caudate (basal ganglia)
Brain − Putamen (basal ganglia)
Thyroid
Brain − Substantia nigra
Brain − Nucleus accumbens (basal ganglia)
Brain − Hippocampus
Brain − Amygdala
Brain − Spinal cord (cervical c−1)
Adipose − Visceral (Omentum)
Adipose − Subcutaneous
Ovary
Vagina
−1.0 −0.5 0.0 0.5 1.0
direction of XCI
(shared cross−tissue SNP XCI ratio correlations)
E
Donor proportion of tissues with switched XCI
0.5
0.4
0.3
0.2
0.1
0.0
mean ratio < .6
.6 <= mean ratio < .7
.7 <= mean ratio < .8
.8 <= mean ratio < .9
Donor mean tissue XCI ratios
F
Proportion of samples with switched XCI
0.25
0.20
0.15
0.10
0.05
0.00
* : p-value <= .05, Fisher's exact test for enrichment of switched XCI direction
Minor Salivary Gland
Liver
Bladder
Brain − Cerebellum
Heart − Left Ventricle
Pancreas
Adrenal Gland
Whole Blood
Artery − Coronary
Esophagus − Mucosa
Vagina
Skin − Not Sun Exposed (Suprapubic)
Ovary
Kidney − Cortex
Skin − Sun Exposed (Lower leg)
Stomach
Uterus
Brain − Caudate (basal ganglia)
Brain − Cortex
Brain − Cerebellar Hemisphere
Colon − Transverse
Spleen
Heart − Atrial Appendage
Brain − Hippocampus
Small Intestine − Terminal Ileum
Thyroid
Lung
Breast − Mammary Tissue
Muscle − Skeletal
Artery − Tibial
Esophagus − Muscularis
Pituitary
Adipose − Subcutaneous
Artery − Aorta
Esophagus − Gastroesophageal Junction
Adipose − Visceral (Omentum)
Nerve − Tibial
Brain − Amygdala
Brain − Anterior cingulate cortex (BA24)
Brain − Frontal Cortex (BA9)
Brain − Hypothalamus
Brain − Nucleus accumbens (basal ganglia)
Brain − Putamen (basal ganglia)
Brain − Spinal cord (cervical c−1)
Brain − Substantia nigra
Cervix − Ectocervix
Cervix − Endocervix
Colon − Sigmoid
Fallopian Tube
Kidney − Medulla

Figure 2.5: ***Individual tissue lineages exhibit increased variance in XCI ratios***
A, Folded allele-specific expression distributions for individual tissues from the 11P81 donor with the aggregated germ layer distributions in the top panel. **B**, Folded allele-specific expression distributions for individual tissues from the 1J1OQ donor with the aggregated germ layer distributions in the top panel. **C**, Pearson correlation distributions calculated from all pairwise comparisons of shared heterozygous SNPs between two tissues for all of donor 11P81's tissues. Positive correlations indicate the same parental direction of XCI, negative correlations indicate opposite parental directions of XCI. **D**, Similar to C, displaying results for donor 1J1OQ's tissues. **E**, Box plots of the per donor proportion of tissues that switched parental XCI directions with donors binned by their mean XCI ratio across tissues. **F**, Bar plot indicating the proportion of donors where the specified tissue switched directions compared to other tissues. Asterisks indicate significance from Fisher's Exact test (FDR corrected p-value <= .05), identifying tissues enriched for switching XCI directions.

2.6.5 Cell population estimate at the time of embryonic epiblast lineage specification

The fact XCI ratios are broadly shared across tissues suggests the initial embryonic XCI ratio determined at the time of inactivation is propagated through development. This is strongly evidenced by the consistency of XCI ratios across the developmentally distant germ layer lineages (Fig. 2.4, Fig. 2.5A-B). Population level variance in adult XCI ratios thus, in part, reflects the sample distribution during XCI, which depends on the number of cells present during inactivation. We derive estimates for the number of cells present at the time of inactivation by modeling XCI ratio variance from tissue-specific ratio distributions across donors (Fig. 2.6A). Using a maximum likelihood approach, we fit estimated models to the tails of the empirical XCI ratio distributions to account for the uncertain unfolded XCI ratio estimates between 0.4 and 0.6 (Fig. 2.6A, see methods). The cell number estimates derived from all well-powered tissues range from 6 to 16 cells (Fig. 2.6B), i.e., approximately within a single cell division, demonstrating a striking degree of similarity in population level XCI ratio variance across the assessed tissues. We model variance in XCI ratios as a random binomial sampling event that is then propagated through development. The consistency in XCI ratios across developmentally distant tissues supports this model, though there are likely additional contributors to the observed variance in XCI ratios, such as genetic variation which might drive allelic selection (Brown and Robinson 2000; Schmidt and Sart 1992) as well as stochastic deviations during development (Sun et al. 2021). In the simplest case, observed variance in XCI ratios is derived from the initial stochasticity of XCI, positing our cell number estimates as lower bounds for the number of cells that must be involved in XCI.

Notably, we sample variance in XCI of tissues derived from the embryonic lineage. If XCI occurs before extraembryonic/embryonic lineage specification, the variance we observe in adult tissues is a combination of the initial variance at the time of XCI and additional sampling variance linked to the lineage specification of the embryonic epiblast. This contextualizes our 6-16 cell number estimate as a potential lower bound for the number of cells present during embryonic epiblast lineage specification in the human embryo.

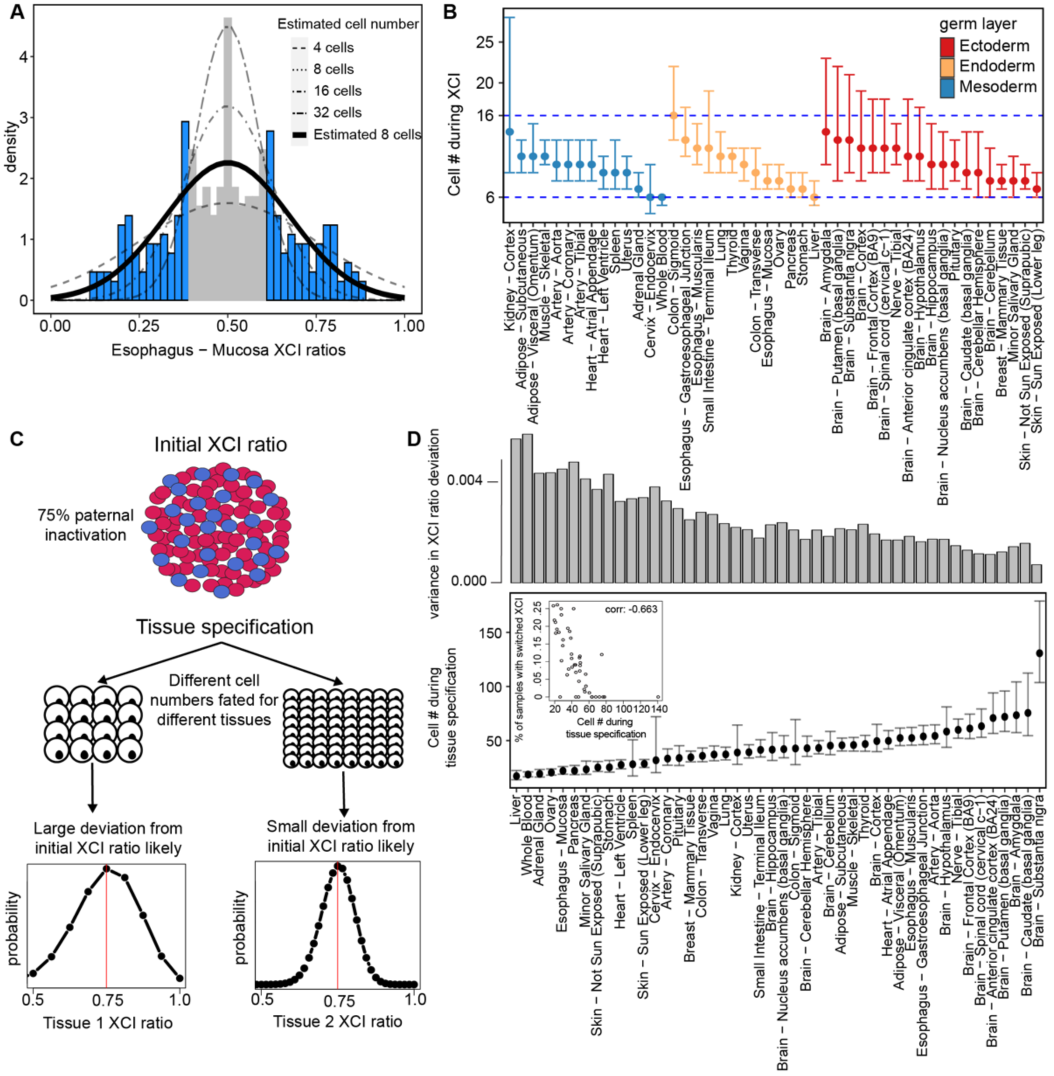
A
Estimated cell number
4 cells
8 cells
16 cells
32 cells
Estimated 8 cells
density
0
1
2
3
4
0.00
0.25
0.50
0.75
1.00
Esophagus – Mucosa XCI ratios
B
germ layer
Ectoderm
Endoderm
Mesoderm
Cell # during XCI
6
16
20
25
Kidney – Cortex
Adipose – Subcutaneous
Adipose – Visceral (Omentum)
Muscle – Skeletal
Artery – Aorta
Artery – Coronary
Artery – Tibial
Heart – Atrial Appendage
Heart – Left Ventricle
Spleen
Uterus
Adrenal Gland
Cervix – Endocervix
Whole Blood
Colon – Sigmoid
Esophagus – Gastroesophageal Junction
Esophagus – Muscularis
Small Intestine – Terminal Ileum
Lung
Thyroid
Vagina
Colon – Transverse
Esophagus – Mucosa
Ovary
Pancreas
Stomach
Liver
Brain – Amygdala
Brain – Putamen (basal ganglia)
Brain – Substantia nigra
Brain – Cortex
Brain – Frontal Cortex (BA9)
Brain – Spinal cord (cervical c–1)
Nerve – Tibial
Brain – Anterior cingulate cortex (BA24)
Brain – Hypothalamus
Brain – Hippocampus
Brain – Nucleus accumbens (basal ganglia)
Pituitary
Brain – Caudate (basal ganglia)
Brain – Cerebellar Hemisphere
Brain – Cerebellum
Breast – Mammary Tissue
Minor Salivary Gland
Skin – Not Sun Exposed (Suprapubic)
Skin – Sun Exposed (Lower leg)
C
Initial XCI ratio
75% paternal inactivation
Tissue specification
Different cell numbers fated for different tissues
Large deviation from initial XCI ratio likely
Small deviation from initial XCI ratio likely
probability
0.5
0.75
1.0
Tissue 1 XCI ratio
Tissue 2 XCI ratio
D
variance in XCI ratio deviation
0.000
0.004
Cell # during tissue specification
50
100
150
% of samples with switched XCI
0
.05
.10
.15
.20
.25
corr: -0.663
20
40
60
80
100
120
140
Cell # during tissue specification
Liver
Whole Blood
Adrenal Gland
Ovary
Esophagus – Mucosa
Pancreas
Minor Salivary Gland
Skin – Not Sun Exposed (Suprapubic)
Stomach
Heart – Left Ventricle
Spleen
Skin – Sun Exposed (Lower leg)
Cervix – Endocervix
Artery – Coronary
Pituitary
Breast – Mammary Tissue
Colon – Transverse
Vagina
Lung
Kidney – Cortex
Uterus
Small Intestine – Terminal Ileum
Brain – Hippocampus
Brain – Nucleus accumbens (basal ganglia)
Colon – Sigmoid
Brain – Cerebellar Hemisphere
Artery – Tibial
Brain – Cerebellum
Adipose – Subcutaneous
Muscle – Skeletal
Thyroid
Brain – Cortex
Heart – Atrial Appendage
Adipose – Visceral (Omentum)
Esophagus – Muscularis
Esophagus – Gastroesophageal Junction
Artery – Aorta
Brain – Hypothalamus
Nerve – Tibial
Brain – Frontal Cortex (BA9)
Brain – Spinal cord (cervical c–1)
Brain – Anterior cingulate cortex (BA24)
Brain – Putamen (basal ganglia)
Brain – Amygdala
Brain – Caudate (basal ganglia)
Brain – Substantia nigra

Figure 2.6: ***XCI and tissue lineage specification can be timed to a pool of cells by exploiting observed variability***

A, Example tissue demonstrating the model for estimating cell numbers at the time of XCI using the population-level variance in XCI ratios. We fit normal distributions, as a continuous approximation of the underlying binomial distribution of XCI ratios, to the tails of tissue-specific XCI ratio distributions (shaded in blue), which accounts for the uncertain 0.40-0.60 unfolded XCI ratio estimates (shaded in grey). **B**, The resulting estimated cell numbers present during XCI derived from the XCI ratio variance of all tissues with at least 10 donors. Error bars are 95% confidence intervals and tissues are grouped by germ layer lineage. **C**, Schematic for our model of tissue lineage specification and the implications for tissue-specific XCI ratios. The XCI ratio of a tissue is dependent on the prior XCI ratio of the embryo and the number of cells selected for that tissue lineage. These two features define the binomial distribution for that tissue's XCI ratio. **D**, Estimated number of cells selected for individual tissue lineage specification of 46 different tissues. Error bars represent 95% confidence intervals. The top bar graph plots the variance in the distribution of tissue XCI ratio deviation from the average XCI ratio of each donor for that tissue. The inset plot compares the estimated number of cells present at the time of tissue specification to the proportion of that tissue's samples that switched parental XCI directions, Pearson correlation -0.663 (p-value < .001).

2.6.6 Cell population estimates at the time of tissue-specific lineage commitment

Tissue-specific lineage commitment can be modeled as a random sampling event from a pool of unspecified progenitor cells. In the context of XCI, the XCI ratio of the newly specified tissue is dependent on the prior XCI ratio of the progenitor pool and the number of cells fated for that tissue and can be modeled as a binomial sampling event (Fig. 2.6C). As such, the GTEx dataset offers a unique opportunity to capture this tissue-specific XCI variance and model the lower bound for the number of cells present at the time of tissue-specific lineage commitment across a broad range of human tissues.

To capture the tissue-specific variance in XCI as it relates to the prior embryonic XCI ratio, we model the deviation of tissue-specific XCI ratios from the average donor XCI ratios for all donors of a given tissue (see methods, Fig. 2.6D, 46 well-powered tissues). Our model follows the logic that tissues with large variation in their deviation from average donor XCI ratios are derived from a smaller pool of cells, a consequence of increased variability due to small sample size effects. On the low end of the estimated cell numbers, we have liver, whole blood, and adrenal tissues with ~20 estimated cells compared to the brain tissues which occupy most of the higher estimated cell numbers, ranging from ~40-140 estimated cells. In line with our model that tissues derived from smaller stem cell pools are subject to increased variability in XCI ratios, we find a strong negative relationship between our estimated tissue lineage-specific cell numbers and the probability of a tissue switching the direction of parental XCI (Fig. 2.6D inset, Pearson correlation: -0.663, p-value < .001). A tissue derived from a small number of cells is more likely to result in a sample of oppositely skewed cells compared to the parental XCI ratio of the unspecified progenitor pool simply through increased sampling variance. Our estimated lineage-specific cell numbers and lineage-specific probability for switching parental XCI are internally consistent with a model of lineage-specific variance in XCI ratios being driven by cell sampling variation at the time of lineage specification.

2.7 Methods

2.7.1 Data and code availability

This paper analyzes existing, publicly available data. Links to access these datasets are listed in the key resources table. The generated allele-specific expression information per sample (variant information removed) and the CIBERSORTx deconvolution results are made available at the FTP site:

http://labshare.cshl.edu/shares/gillislab/people/werner/werner_et_al_Dev_Cell_2022 /data.

Descriptions of the data are available at

github.com/JonathanMWerner/human_cross_tissue_XCI

All original code has been deposited at figshare (DOI: 10.6084/m9.figshare.20216816) and at Github (github.com/JonathanMWerner/human_cross_tissue_XCI) and is publicly available as of the date of publication.

2.7.2 GTEx and EN-TEx data

Fastq files for all female donors from the GTEx project v7 release (Lonsdale et al. 2013) were obtained from dbGaP accession number phs000424.vN.pN. BAM files for additional female samples from the v8 release were obtained from the associated AnVIL repository (gtexportal.org/home/protectedDataAcccess). All GTEx v7 data files can also be accessed in the GTEx v8 AnVIL repository. Phased expression data from the EN-TEx project (Rozowsky et al. 2021) were obtained in collaboration with the ENCODE consortium. EN-TEx data is available on the online portal. Expression data and annotations for the GTEx single nucleus RNA-sequencing data were obtained from the GTEx data portal.

2.7.3 RNA-seq alignment and SNP identification

For aligning RNA-sequencing data, the GRCh38.p7 human reference genome using GENCODE v.25 (Frankish et al. 2021) annotations was generated with STAR v2.4.2a (Dobin et al. 2013) and data was aligned with STAR v2.4.2a or STAR v2.5.2b. STAR was run using default parameters with per sample 2-pass mapping. BAM files for the additional GTEx v8 samples (originally aligned to GRCh38.p10 with GENCODE v.26 annotations) were sorted using samtools v1.9 (Li et al. 2009) and converted to fastq files using bedtools v.2.26.0 (Quinlan and Hall 2010). For each sample, alignment to the X-chromosome was extracted using samtools and passed to GATK (McKenna et al. 2010) for SNP identification. Using GATK v.4.1.3.0 and following the best practices workflow for RNAseq short variant discovery (GATK best practices), we utilized the following pipeline of GATK tools using default parameters unless otherwise stated: AddorReplaceReadGroups -> MarkDuplicates ->

SplitNCigarReads -> HaplotypeCaller (-stand-call-conf 0.0) -> SelectVariants (-select-type SNP) -> VariantFiltration. The following filters were used in VariantFiltration to set flags for downstream filtering: QD < 2.0, QUAL < 30.0, SOR > 3.0, FS > 60.0, MQ < 40.0, MQRankSum < -12.5, and ReadPosRankSum < -8.0. These filters were determined from GATK recommendations and empirical evaluation of the identified SNPs' metrics.

2.7.4 SNP quality control

SNPs identified through GATK were further filtered on various metrics to increase confidence in SNPs identified from RNA-sequencing data and ensure well-powered SNPs for allele-specific expression analysis. The resulting .vcf files from GATK were filtered to only contain SNPs present within dbSNP (Sherry et al. 2001). The remaining SNPs were filtered to be heterozygous with 2 identified alleles and at least 10 reads mapped to each allele for a minimum threshold of 20 reads per SNP. Additionally, SNPs were required to pass the SOR, FS, and ReadPosRankSum filters set in the GATK pipeline. Only SNPs located within annotated genes (excluding the PAR regions of the X-chromosome) were considered and in the case of multiple identified SNPs in the same gene for a sample, the SNP with the highest total read count was taken as the max-powered representative for that gene. SNPs with a total read count above 3000 were excluded as they demonstrated a uniform distribution of allelic expression.

2.7.5 Gene filtering (reference bias and XCI escape)

From the observation of a heavy tail towards allelic expression in the reference direction across all called SNPs in the GTEx dataset, we compiled gene specific distributions of allelic expression to determine if a select few genes/SNPs were at fault. The majority of genes demonstrated distributions of relative allelic expression centered around 0.5 with several considerable exceptions, some genes exhibited bimodal or extremely biased distributions. We excluded genes that failed the dip test for unimodality as well as the top and bottom 5% of genes ranked by the deviation of their mean reference expression ratio from 0.5. Additionally, we excluded genes previously annotated to constitutively escape XCI (Tukiainen et al. 2017). In total, we end up with well-powered SNPs from 542 genes along the X-chromosome for modeling XCI ratios.

2.7.6 Folded normal model for estimating XCI ratios

We aggregate the allelic expression imbalance of the X-chromosome over both alleles by folding the reference allelic expression ratios about 0.5 (Fig 2.2A-B). To obtain our XCI ratio estimates we fit a folded normal distribution to the folded reference allelic expression ratios of each sample, using the maximum log likelihood estimate as the estimated XCI ratio. Theoretically, the captured bulk allelic expression for a heterozygous X-linked SNP follows a binomial distribution characterized by the read depth of the SNP and the XCI ratio of the sample. Without phasing information, the allelic expression of heterozygous X-linked SNPs can be characterized by the folded-binomial model (Gart 1970; Urbakh 1967). Since SNPs vary in read depth and various biological factors (e.g. eQTLs) are not accounted for in the binomial model, we take the folded normal model as a continuous approximation. We require samples to have XCI ratio estimates derived from at least 10 filtered SNPs for downstream analysis, resulting in 4659 samples with a mean of 56 well-powered SNPs per sample (Fig. S2.1). Additionally, we calculate 95% confidence intervals (CI) for each XCI ratio estimate via a nonparametric bootstrap percentile approach (n = 200), excluding XCI ratio estimates with a CI width >= .15 from downstream analysis. For donors with multiple samples for the same tissue, we average the XCI ratio estimates together, duplicated tissue samples have minor differences in estimated XCI ratios (mean difference in XCI ratios for duplicate tissue samples: 0.018 +- 0.023 SD).

2.7.7 Modeling read sampling error when estimating XCI ratios

The sampled allelic reads for any expressed heterozygous loci will follow a binomial distribution defined by the total number of reads sampled (n) and the probability for allelic expression (p). For a given GTEx sample, we define SNP-specific binomial distributions as Binomial(n = total number of reads, p = sampled reference allelic expression ratio). For each individual GTEx tissue sample, we randomly sample a single instance from each SNP-specific binomial distribution to simulate SNP expression ratios with noise from allelic read sampling. We estimate the XCI ratio using the folded normal model on the simulated SNP expression ratios and repeat the simulation 50 times to generate a distribution of estimated tissue XCI ratios. We compute the root mean squared error of the simulated tissue XCI ratios about the original estimated tissue XCI ratio. We repeat the entire analysis with a percent reduction in each SNP's total read count (10%, 20%, 30%, etc.) to model variance in our estimated XCI ratios as read depth decreases.

2.7.8 Gene-tissue XCI ratio correlations

To test individual gene's propensity to follow the aggregate chromosomal XCI ratio, we calculate Pearson correlations between a gene's reference allelic expression ratio and the estimated XCI ratio leaving out that gene for all samples the gene is detected. We calculate these correlations for each of the 542 filtered genes described above and for 45 previously annotated constitutively escape genes detected in our dataset. We only consider genes detected in at least 30 samples and with an FDR corrected (Benjamini-Hochberg) correlation p-value <= .05 determined by a permutation test (n = 10000) for further investigation of escape status, resulting in 380 putative inactive genes and the 45 previously annotated escape genes.

2.7.9 Testing for escape from XCI

To detect escape genes, it is necessary to compare against genes that undergo complete inactivation and do not escape. After stratifying by mean expression, we reason the genes most likely to undergo complete inactivation are genes with high gene-tissue XCI ratio correlations within each expression bin (Fig. 2.3C). Accordingly, we take the top 50% of putative inactive genes within each bin to define the null distribution of allelic expression under the hypothesis of complete inactivation (191 genes). The remaining 189 putative inactive genes and the 45 known escape genes comprise our test set. We reason a gene that escapes XCI will be biased for balanced biallelic expression regardless of the XCI ratio of the tissue. Using only tissues with an estimated XCI ratio >= 0.70, we compute the deviation from 0.5 (balanced allelic expression) for all inactive genes and the test gene. We rank the gene deviations and calculate the empirical p-value as the rank of the test gene divided by the total number of ranks i.e. the number of null inactive genes + 1 (Fig. S2.2). We only consider empirical p-values derived from samples with at least 20 null inactive genes detected. Additionally, we only consider test genes with at least 50 empirical p-values. For each remaining test gene, we aggregate the distribution of empirical p-values using Fisher's method and apply an FDR correction (Benjamini-Hochberg) to the resulting meta-analytic p-values. We use a threshold of meta-analytic p-value < .001 to call significance for escape. For Fisher's method, under the null hypothesis, the log sum of all p-values follows a chi-squared distribution with 2k degrees of freedom, where k is the number of independent tests being combined. We use R's pchisq function to compute the meta-analytic p-value for the following test statistic:

$$X_{2k}^{2} \sim -2\sum_{i=1}^{k}\log(p_i).$$

2.7.10 Tissue XCI ratio predicting donor XCI ratio

For the donors that contribute to a given tissue, we calculate the mean XCI ratio across all other tissues for each donor and use that mean as an approximation for the true XCI ratio for each donor. We classify donors as low/high XCI ratio donors if they have a mean XCI ratio greater than or equal to various thresholds (0.65, 0.7, 0.75). We calculate the AUROC of a given tissue's XCI ratio predicting the low/high donors via the Mann-Whitney U test statistic where

$$AUC_{tissue} = \frac{U}{n_{high\ donors} n_{low\ donors}}.$$

2.7.11 Cross-tissue XCI ratio correlations

For all pairwise combinations of the 49 tissues present within the GTEx dataset, we take the subset of donors that contribute both tissues for a given comparison and calculate the Pearson correlation for the folded XCI ratio of the tissues. Figure 2.4c1-c2 depicts only the correlation values derived with a sample size of at least 20 donors and an FDR corrected (Benjamini-Hochberg) p-value <= .05 derived from a permutation test (n = 10000). Supplemental Figure 2.6 depicts all computed correlations regardless of sample size or p-value.

2.7.12 CIBERSORTx deconvolution and germ layer-specific marker identification

CIBERSORTx (https://cibersortx.stanford.edu, (Newman et al. 2019)) was run using the recommended settings following the "Build a signature matrix file from single-cell RNA sequencing data" and "Impute cell fractions" tutorials, batch correction was enabled when imputing cell fractions. Briefly, the annotated single-cell RNA sequencing data from GTEx is used to build a signature matrix that identifies genes that define the annotated cell types. This signature matrix is used to impute the cell type composition of bulk RNA sequencing samples. We extract germ layer-specific marker genes from the signature matrices identified from CIBERSORTx, classifying a gene as a germ layer marker if it is a gene that identifies cell types exclusively from a single germ layer. Our annotated germ layer markers, the cell types they define, and the tissue they are derived from are available in Supplementary Table 2.2. The signature matrices and imputed cell types per tissue with associated statistics from CIBERSORTx are made available on the FTP site http://labshare.cshl.edu/shares/gillislab/people/werner/werner_et_al_Dev_Cell_2022 /data.

2.7.13 Inference on direction of XCI ratios

To infer the direction of XCI ratios from unphased data, we look at allelic expression of heterozygous SNPs captured in multiple tissues for an individual donor. The reference allele of a heterozygous SNP captured in two different tissues of a single donor represents the same parental X-allele in both tissues. If the direction of XCI is the same for both tissues, the heterozygous SNP is expected to exhibit the same degree of reference allelic expression across the two tissues (positive correlation). If the direction of XCI is different, reference allelic expression will be inverted for one of the tissues resulting in a negative correlation. For each donor, for all pairwise combinations of their donated tissues with XCI ratios >= 0.6, we calculate Pearson correlations for unfolded reference allelic expression ratios using only SNPs detected in both tissues (Fig. 2.5). We only use SNPs that are within the previously filtered 542 genes described above and only consider correlations derived from tissue comparisons with at least 30 shared SNPs. Using positive or negative correlations as a readout for switched XCI direction between tissues, we perform Fisher’s exact test with a Benjamini-Hochberg correction to identify any tissue significantly enriched for switching XCI directions. We use the hypergeometric distribution to calculate raw p-values for Fisher’s Exact Test. For a given tissue, we input the number of times that tissue switched XCI directions minus 1, the total number of switched XCI cases across all tissues, the total number of non-switched XCI cases across all tissues, and the sample size for the given tissue.

2.7.14 Evaluating XCI cell number estimates

XCI is a binomial sampling event defined by the number of cells present during inactivation and the equal probability of inactivation between the alleles Binomial(N = # of cells, p = 0.5). As such, the variance in XCI ratios within a population is directly linked to the number of cells present during XCI. We derive estimates for the number of cells present during XCI by fitting a normal model to tissue-specific XCI skew distributions as a smoothened estimate for the underlying binomial distribution. We take the theoretical variance from the binomial model as the variance for the normal approximation.

$var_{XCI} = var\left(\frac{Binomial(N,p,q)}{N}\right) = \frac{pq}{N} = \frac{.5(1-.5)}{N_{embryo}}$, where p,q = probability of allelic inactivation.

For a range of cell numbers (N = 2:50), we select the normal model with minimum error between its CDF and the empirical XCI ratio CDF of a given tissue for the tails of the distribution (XCI ratio <= 0.4 and XCI ratio >= 0.6). This accounts for the uncertain folded 0.5 – 0.6 XCI ratios estimates in the unfolded space. We calculate 95% CIs for each estimated cell number via a nonparametric bootstrap percentile approach (n = 2000). We only consider cell number estimates from tissues with at least 10 donors.

2.7.15 Evaluating tissue-specific lineage cell number estimates

We model tissue-specific lineage specification as a cell sampling event from a large pool of cells. As such, the XCI ratio of a tissue will follow a binomial model defined by the number of cells fated for that tissue and the XCI ratio of the embryo (Fig. 2.6c).

$$XCI_{tissue} \sim \frac{Binomial(N, p, q)}{N} = \frac{Binomial(N_{tissue},\ XCI_{embryo},\ 1 - XCI_{embryo})}{N_{tissue}}$$

$$var_{XCI_{tissue}} = var\left(\frac{Binomial(N, p, q)}{N}\right) = \frac{pq}{N} = \frac{XCI_{embryo}(1 - XCI_{embryo})}{N_{tissue}}$$

$$SD_{XCI_{tissue}} = \sqrt{\frac{XCI_{embryo}(1 - XCI_{embryo})}{N_{tissue}}}$$

For a given tissue, across donors with variable XCI ratios (XCI_{embryo}) the variation in the tissue XCI ratio is defined by the constant N_{tissue}, the number of cells fated for that tissue. To estimate this constant, we calculate z-scores for each tissue-donor pair of a given tissue using the mean XCI ratio of all other tissues for each donor as an approximation for the XCI_{embryo}.

$$Z_{tissue} = \frac{XCI_{tissue} - XCI_{embryo}}{SD_{tissue}} = \frac{XCI_{tissue} - XCI_{embryo}}{\sqrt{XCI_{embryo}(1 - XCI_{embryo})}}\sqrt{N_{tissue}}$$

$$= t_{tissue}\sqrt{N_{tissue}}$$

As the standard deviation of a distribution of z-scores is 1, we solve for N_{tissue}:

$SD(Z) = \sqrt{\frac{1}{m-1}\sum_{i=1}^{m}(Z_i - \bar{Z})^2} = 1$, where m = number of donors for a given tissue

$$N_{tissue} = \frac{m-1}{\sum_{i=1}^{m}(t_i - \bar{t})^2}$$

We calculate 95% CIs for each N_{tissue} via a nonparametric bootstrap percentile approach (n = 2000) using the t_{tissue} distribution. We require a tissue to have at least 10 donors in order to calculate N_{tissue}.

2.7.16 Data analysis and visualization

All analysis was conducted in R version 4.0.5 (R Core Team 2021). Graphs were generated using the ggplot2 (Wickham 2016), ComplexHeatmap (Gu et al. 2016), karyoploteR (Gel and Serra 2017), and base R packages.

2.7.17 Quantification and statistical analysis

When correcting p-values, we use the Benjamini-Hochberg procedure implemented by R's p.adjust function with “method = BH” parameter. Significance is determined with p-value <= 0.05 unless otherwise stated. We use the R dip.test function from the diptest package to perform Hartigan’s dip test of unimodality. For Fisher’s method of aggregating p-values, we use the R function pchisq with ‘lower.tail = FALSE’ parameter to compute the meta-analytic p-value from the calculated chi-square test statistic. All confidence intervals are computed using a nonparametric bootstrap percentile approach, where the underlying data is sampled with replacement to generate a bootstrapped distribution of the variable in question (tissue XCI ratio estimates, cell number estimates). The 95% confidence interval is defined by the 2.5th and 97.5th percentile of the bootstrapped distribution. We determine if tissues are enriched for switching parental XCI directions using the hypergeometric implementation of Fisher’s Exact Test, using R’s phyper function. When fitting normal distributions to tissue XCI ratio distributions, we use the R quantile function with parameter “type = 1” to compute the empirical CDF and the R qnorm function to compute the theoretical normal CDF. For any given correlation calculated, we permute the underlying data to get a null distribution of correlations under the hypothesis of independence, using R’s cor function with “method =

pearson” parameter. We derive a raw p-value for the original correlation value from the empirical null distribution of correlations (permutation test). In the analyses where we generate many correlations, we apply a Benjamini-Hochberg FDR correction to the associated distribution of raw p-values to call significance, using a threshold of p-value <= 0.05.

2.8 Supplemental Figures

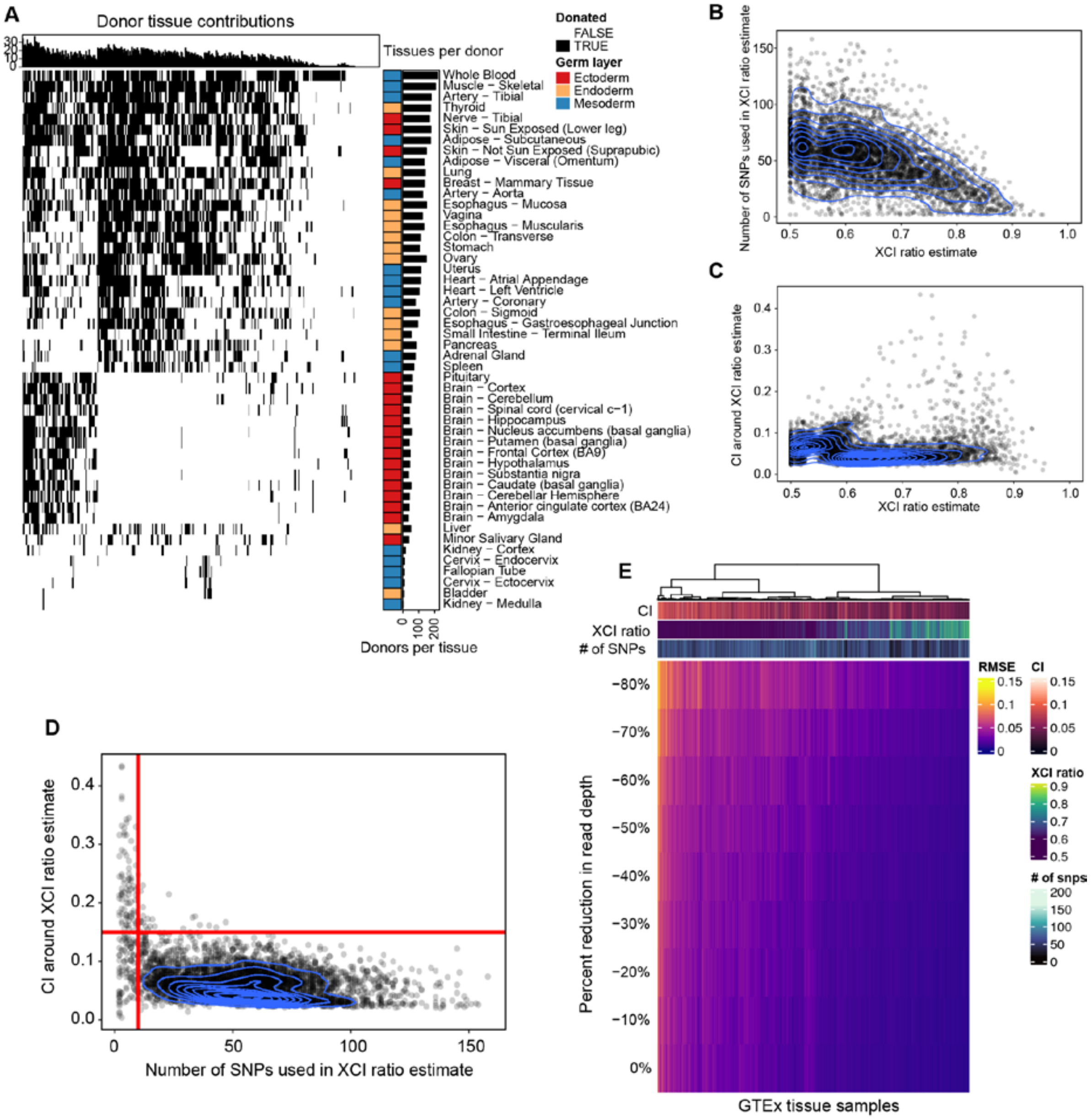


Figure S2.1: ***Estimating robust XCI ratios from GTEx tissue samples. Related to Figure 2.2***

A, Binary heatmap of female donor tissue contributions in the GTEx dataset for samples that pass our quality control filters. Data from cell lines was excluded in the final analysis.
B, Scatter plot with 2d density overlay of all XCI ratio estimates for 5046 GTEx samples and the number of filtered heterozygous SNPs used to estimate the sample XCI ratio.
C, Scatter plot with 2d density overlay of all XCI ratio estimates for 5046 GTEx samples and the width of the 95% confidence interval around the XCI ratio estimate (bootstrap sampling, n = 200).
D, Scatter plot with 2d density overlay of the number of filtered heterozygous SNPs used to estimate the sample XCI ratio and the width of the 95% confidence interval around the XCI ratio estimate. Red lines indicate thresholds for XCI ratio estimate filtering, requiring >= 10 heterozygous SNPs and a CI width < 0.15.
E, Matrix of root mean squared error for estimated XCI ratios per GTEx sample as read depth per SNP is gradually reduced. Tissue sample annotations for the confidence interval about the original XCI ratio estimate (CI), the original XCI ratio estimate (XCI ratio), and the number of SNPs used to estimate the original XCI ratio estimate (# of snps) are provided as column annotations.

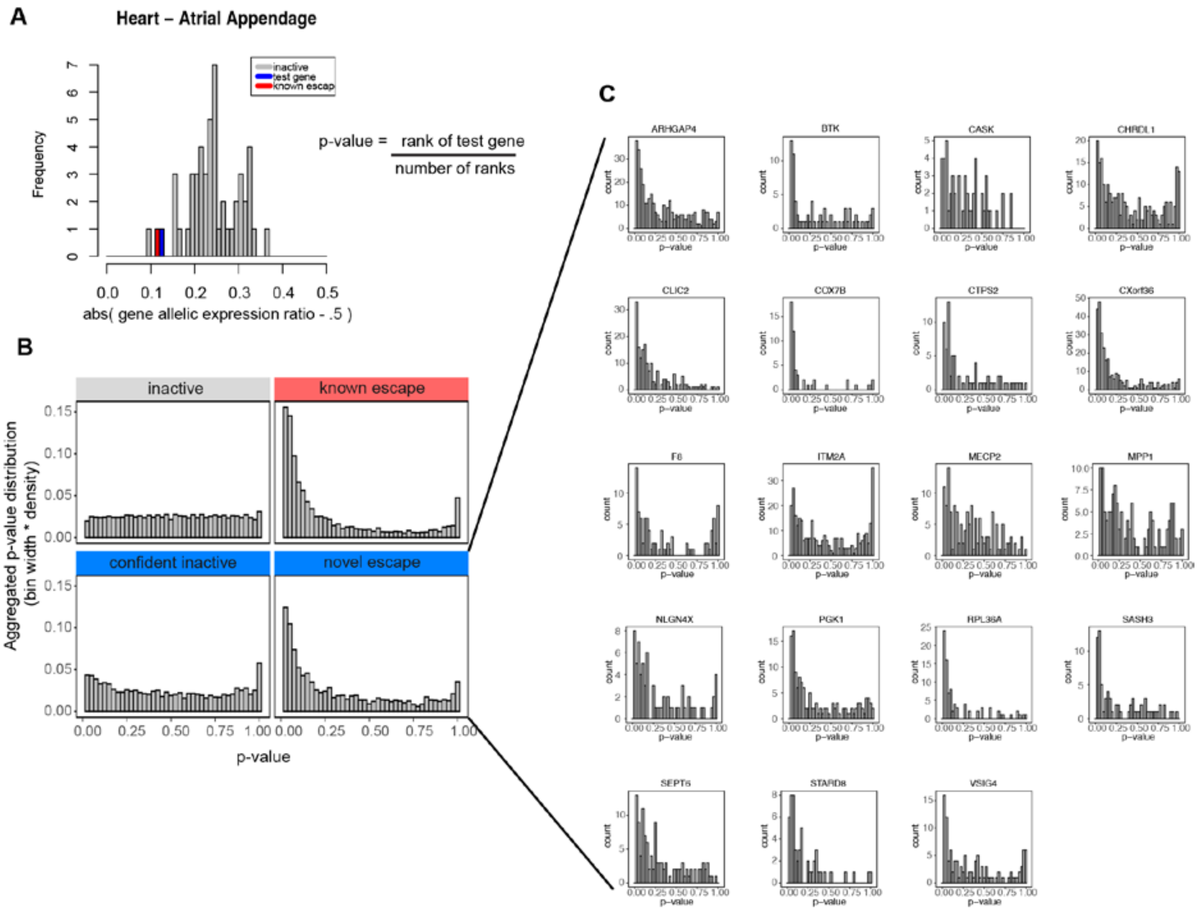


Figure S2.2: ***XCI escape genes exhibit balanced allelic expression in skewed XCI tissues. Related to Figure 2.3***

A, Histogram of gene reference allelic expression ratio deviations from 0.5 for a sample with an estimated XCI ratio >= 0.70. An example known escape gene in the sample is colored red, an example putative inactive gene is colored blue, and the inactive genes are colored in grey. After ranking the allelic expression ratio deviations, the empirical p-value for the given test gene is calculated as the rank of the test gene divided by the total number of ranks, i.e. the number of inactive genes plus one.
B, Central histograms are the same plots as in Figure 2.3E.
C, empirical p-value distributions for the 19 genes that we classify as novel escape genes, each gene has a FDR corrected meta-analytic p-value (Fisher's method) < .001.

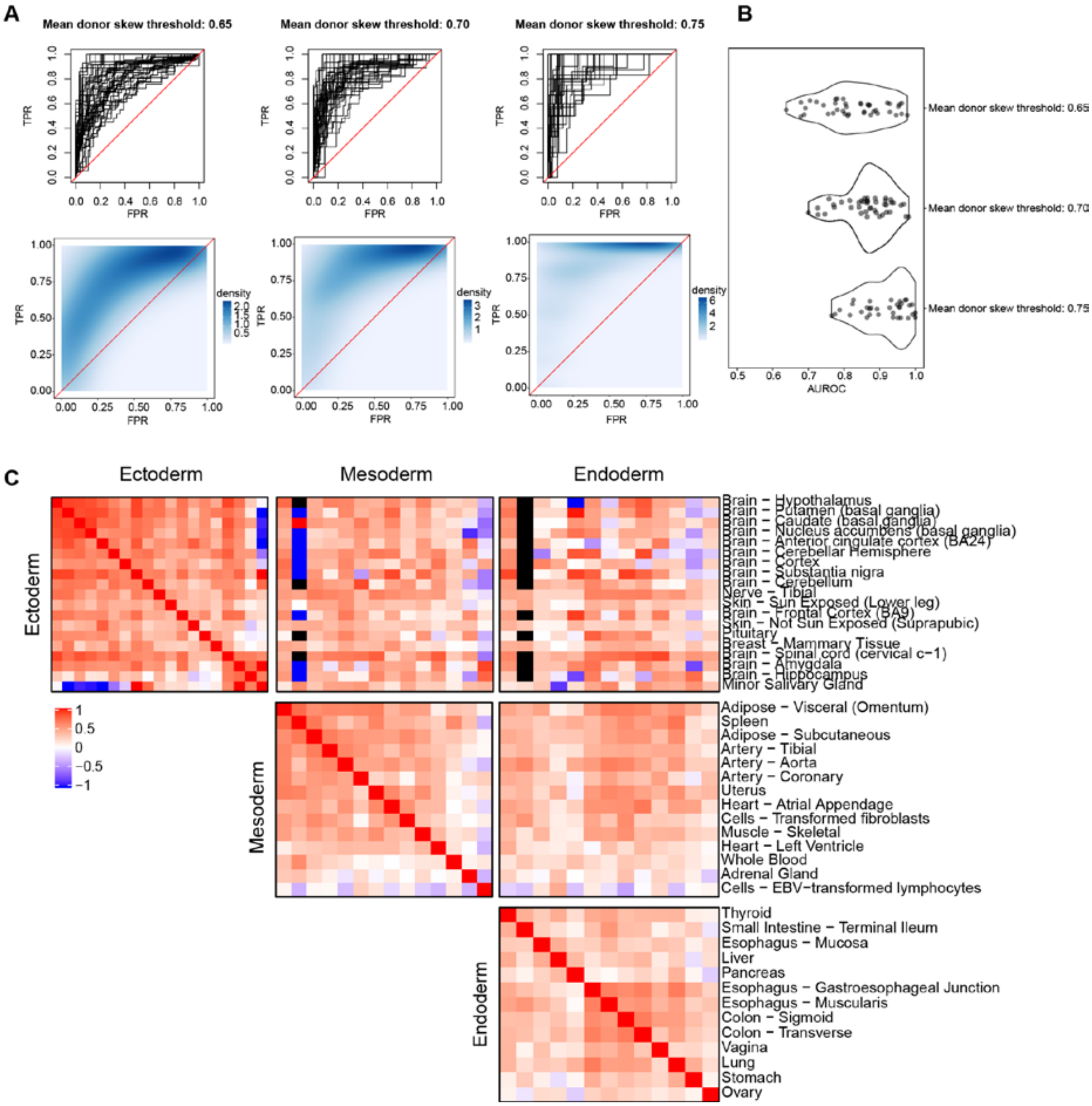


Figure S2.3: ***All tissues strongly predict skewed donors and are correlated in XCI ratios. Related to Figure 2.4***

A, ROC curves for individual tissue XCI ratios predicting skewed donors at various thresholds for classifying skewed donors (top row). 2d density estimations across all tissue ROC curves (bottom row).
B, AUROC distributions at each skewed donor threshold.
C, All pairwise tissue-tissue XCI ratio correlations regardless of sample size or significance, grouped by germ layer lineage. The global trend is a positive correlation.

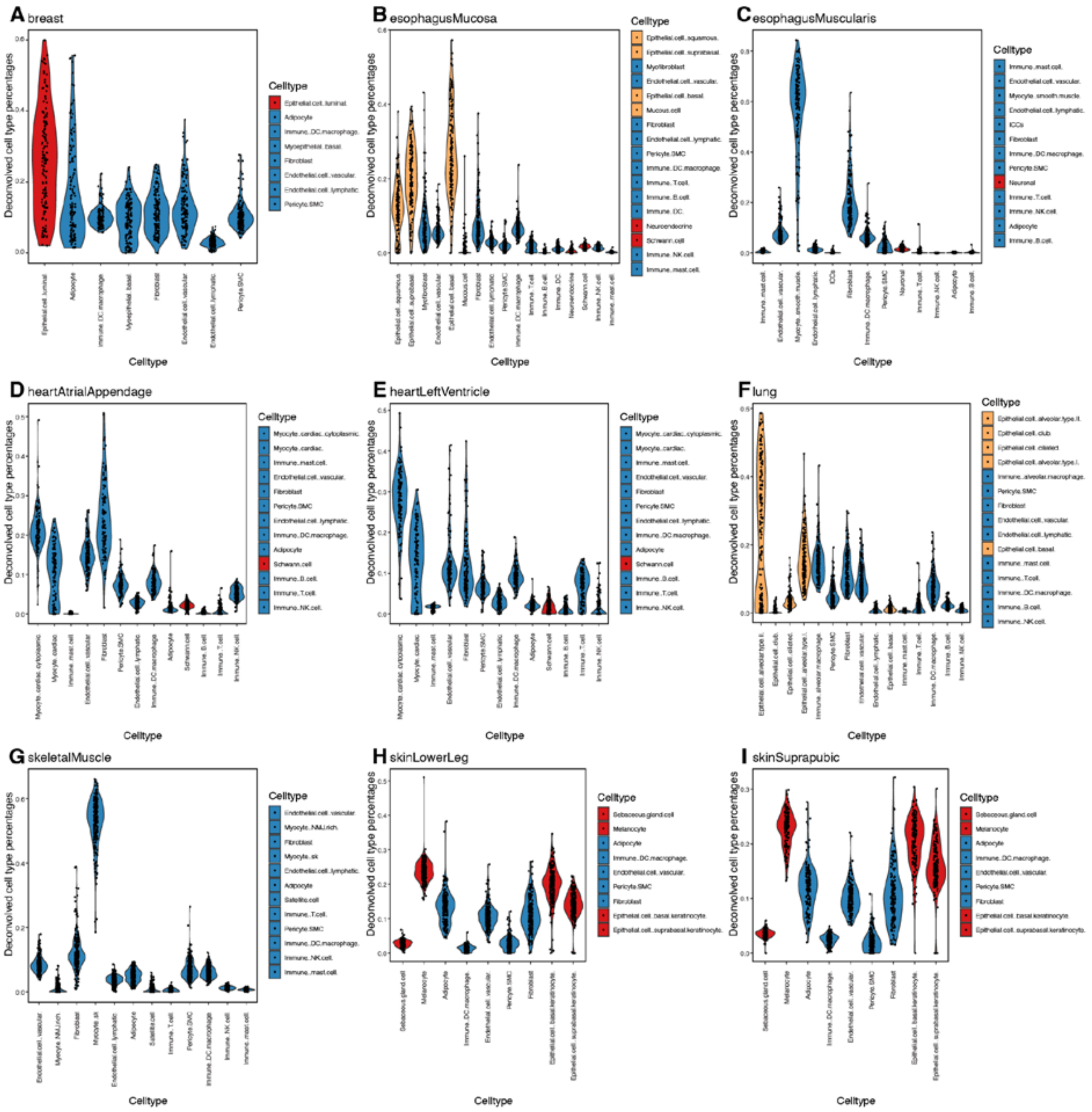

A breast
Deconvolved cell type percentages
0.0
0.2
0.4
0.6
Celltype
Epithelial.cell..luminal.
Adipocyte
Immune..DC.macrophage.
Myoepithelial..basal.
Fibroblast
Endothelial.cell..vascular.
Endothelial.cell..lymphatic.
Pericyte.SMC
B esophagusMucosa
Deconvolved cell type percentages
Celltype
Epithelial.cell..squamous.
Epithelial.cell..suprabasal.
Myofibroblast
Endothelial.cell..vascular.
Epithelial.cell..basal.
Mucous.cell
Fibroblast
Endothelial.cell..lymphatic.
Pericyte.SMC
Immune..DC.macrophage.
Immune..T.cell.
Immune..B.cell.
Immune..DC.
Neuroendocrine
Schwann.cell
Immune..NK.cell
Immune..mast.cell.
C esophagusMuscularis
Deconvolved cell type percentages
0.0
0.2
0.4
0.6
0.8
Celltype
Immune..mast.cell.
Endothelial.cell..vascular.
Myocyte..smooth.muscle.
Endothelial.cell..lymphatic.
ICCs
Fibroblast
Immune..DC.macrophage.
Pericyte.SMC
Neuronal
Immune..T.cell.
Immune..NK.cell.
Adipocyte
Immune..B.cell.
D heartAtrialAppendage
Deconvolved cell type percentages
0.0
0.1
0.2
0.3
0.4
0.5
Celltype
Myocyte..cardiac..cytoplasmic.
Myocyte..cardiac.
Immune..mast.cell.
Endothelial.cell..vascular.
Fibroblast
Pericyte.SMC
Endothelial.cell..lymphatic.
Immune..DC.macrophage.
Adipocyte
Schwann.cell
Immune..B.cell.
Immune..T.cell.
Immune..NK.cell.
E heartLeftVentricle
Deconvolved cell type percentages
0.0
0.1
0.2
0.3
0.4
0.5
Celltype
Myocyte..cardiac..cytoplasmic.
Myocyte..cardiac.
Immune..mast.cell.
Endothelial.cell..vascular.
Fibroblast
Pericyte.SMC
Endothelial.cell..lymphatic.
Immune..DC.macrophage.
Adipocyte
Schwann.cell
Immune..B.cell.
Immune..T.cell.
Immune..NK.cell.
F lung
Deconvolved cell type percentages
0.0
0.2
0.4
0.6
Celltype
Epithelial.cell..alveolar.type.II.
Epithelial.cell..club.
Epithelial.cell..ciliated.
Epithelial.cell..alveolar.type.I.
Immune..alveolar.macrophage.
Pericyte.SMC
Fibroblast
Endothelial.cell..vascular.
Endothelial.cell..lymphatic.
Epithelial.cell..basal.
Immune..mast.cell.
Immune..T.cell.
Immune..DC.macrophage.
Immune..B.cell.
Immune..NK.cell.
G skeletalMuscle
Deconvolved cell type percentages
0.0
0.2
0.4
0.6
Celltype
Endothelial.cell..vascular.
Myocyte..NMJ.rich.
Fibroblast
Myocyte..sk
Endothelial.cell..lymphatic.
Adipocyte
Satellite.cell
Immune..T.cell.
Pericyte.SMC
Immune..DC.macrophage.
Immune..NK.cell.
Immune..mast.cell.
H skinLowerLeg
Deconvolved cell type percentages
0.0
0.1
0.2
0.3
0.4
0.5
Celltype
Sebaceous.gland.cell
Melanocyte
Adipocyte
Immune..DC.macrophage.
Endothelial.cell..vascular.
Pericyte.SMC
Fibroblast
Epithelial.cell..basal.keratinocyte.
Epithelial.cell..suprabasal.keratinocyte.
I skinSuprapubic
Deconvolved cell type percentages
0.0
0.1
0.2
0.3
Celltype
Sebaceous.gland.cell
Melanocyte
Adipocyte
Immune..DC.macrophage.
Endothelial.cell..vascular.
Pericyte.SMC
Fibroblast
Epithelial.cell..basal.keratinocyte.
Epithelial.cell..suprabasal.keratinocyte.

Figure S2.4: ***Bulk tissue samples represent a mix of germ layer lineages. Related to Figure 2.4***

A, Violin plots of the deconvolved cell type percentages across all bulk breast tissue samples.
B, Violin plots of the deconvolved cell type percentages across all bulk esophagus mucosa tissue samples.
C, Violin plots of the deconvolved cell type percentages across all bulk esophagus muscularis tissue samples.
D, Violin plots of the deconvolved cell type percentages across all bulk heart atrial appendage tissue samples.
E, Violin plots of the deconvolved cell type percentages across all bulk heart left ventricle tissue samples.
F, Violin plots of the deconvolved cell type percentages across all lung breast tissue samples.
G, Violin plots of the deconvolved cell type percentages across all skeletal muscle breast tissue samples.
H, Violin plots of the deconvolved cell type percentages across all skin lower leg breast tissue samples.
I, Violin plots of the deconvolved cell type percentages across all bulk skin suprapubic tissue samples.
Cell types are color coded according to their developmental germ layer origin: Ectoderm (red), Endoderm (yellow), Mesoderm (blue)

2.9. Chapter 2 Summary:

In this work, we model XCI ratios across numerous human tissue lineages within and across individuals, revealing XCI ratios are shared across all three germ-layer lineages. This demonstrates that XCI is completed prior to any tissue specification within the human embryo and the stochastically determined XCI ratio at the time of inactivation is propagated through development to all tissues. We estimate that between 6-16 cells must have been present within the embryonic epiblast at the time of inactivation to explain the observed degree of population XCI ratio variance. We also exploit tissue-specific XCI ratio variability to estimate the number of cells that must have been present during tissue lineage specification, reaching estimates ranging from 20-140 cells. In summary, the analysis of variance in XCI ratios across tissues and individuals is informative for inferring characteristics of early developmental lineage decisions in humans.

3. Population variability of X-chromosome inactivation across 9 mammalian species

3.1. Author contributions and Acknowledgements

J.G. conceived the project. J.M.W. and J.G. designed the experiments and wrote the manuscript. J.M.W. performed the experiments. J.H. and J.M.W performed data management and data processing.

J.G., J.M.W., and J.H. were supported by NIH grants R01MH113005. We thank all members of the Gillis lab and particularly John Lee for assisting in some of the initial data downloading.

3.2. Results summary

We apply our model for estimating XCI ratios from reference aligned bulk RNA-seq across data from 9 mammalian species, utilizing publicly available data from the Sequencing Read Archive (SRA). This work extends beyond previous analyses restricted to human and mouse data, providing a broader evolutionary context for a fundamental feature of mammalian development. We reveal that population variability in XCI ratios is a conserved characteristic of XCI and can be explained through models of embryonic stochasticity rather than genetic factors consistently across all species. In total, after extensive filtering for reference bias, global allelic imbalances, and the number of well-powered heterozygous SNPs, we obtain 130 Macaca, 328 Rat, 624 Pig, 383 Goat, 275 Horse, 731 Sheep, 1328 Cow, and 269 Dog samples with 4877 Human samples from the GTEx dataset. Samples that exhibited consistent allelic imbalances in aggregate on 2 autosomes (global allelic imbalances) were excluded due to presumed effects outside of XCI influencing allele-specific expression. We compile population distributions of XCI ratios and reveal species vary substantially in population XCI ratio variability. We demonstrate that models of embryonic XCI stochasticity explain the observed population XCI variability exceptionally well across species, estimating the number of cells that must have been present in the embryonic epiblast to produce the observed population variability. We then quantify the relationship between XCI ratios and X-linked heterozygosity, revealing chromosome-wide genetic variability has no association with XCI ratios. We additionally quantify the relationship between individual variants and XCI ratios, identifying only a select few variants in each species present in low frequencies with modest associations. Taken together, our results demonstrate a pervasive lack of genetic associations with XCI ratios across mammalian species. Instead, models of stochasticity offer a more general explanation for population XCI ratio variability across mammals. In conclusion, assessments of population XCI ratio variability reveal the population-scale consequences of a conserved stochastic feature of mammalian development.

3.3. Introduction

X-chromosome inactivation (XCI) is an early embryonic milestone every female mammalian embryo must achieve for successful development(Lyon 1961; Migeon 2016; Okamoto et al. 2011). XCI evolved to correct the genetic imbalance resulting from the presence of two X-chromosomes in females compared to the single X-chromosome in male mammals(Ohno 1966). While the exact timing can vary across species(Lyon 1972), XCI typically initiates during preimplantation embryonic development(van den Berg et al. 2009). During this process, one of the two X-alleles in each female cell is independently, randomly, and permanently chosen for transcriptional silencing to match the single X-allele in male embryos(Lyon 1961; Evans et al. 1965; Wu et al. 2014; Mutzel et al. 2019). The inactivated X-allele is inherited through cell divisions, propagating the random choice of allelic inactivation down each cell's subsequent lineage. This produces whole-body mosaicism for allelic X-chromosome expression in each adult mammalian female that originates from the very first days/weeks of embryonic development(Migeon 2013).

In humans, both X-alleles have an equal probability of being inactivated; however, population variability in XCI has been widely observed among adult female populations with individuals ranging from balanced to highly skewed allelic inactivation(Amos-Landgraf et al. 2006; Shvetsova et al. 2019). The allelic XCI ratio of an individual becomes highly consequential in the presence of X-linked disease variants, where the allelic-direction and magnitude of XCI can either confer protection or contribute to disease phenotypes(Migeon 2013; Fang et al. 2021). The interplay between variability in XCI ratios and X-linked genetic disorders has prompted extensive research on the underlying factors that contribute to population variability in XCI, primarily restricted to mouse and human data. Well-supported sources of XCI variability within a population include the inherent stochasticity of XCI(Shvetsova et al. 2019) and various genetic factors(Migeon 1998; Plenge et al. 1997; Belmont 1996), though their relative contributions are widely debated(Brown and Robinson 2000). The limited cross-species data on population XCI variability make it difficult to distinguish between generalizable features of XCI or species-specific mechanisms. For example, copious evidence exists for a general genetic basis of XCI variability across lab mouse strains(Cattanach and Isaacson 1965; Simmler et al. 1993; Sun et al. 2021), prompting hypotheses the same holds true in humans(Peeters et al. 2016). However, evidence for general genetic determinants of XCI in humans is lacking(Brown and Robinson 2000; Peeters et al. 2016; Bolduc et al. 2008), albeit more difficult to capture and assess compared to genetically controlled model organisms. Expanding assessments of population XCI variability across mammalian species stands to elucidate generalizable principles of XCI variability, where models of stochasticity or genetic factors can be tested in the face of evolution.

Considering first a stochastic model for XCI variability, each cell within an embryo at the time of XCI independently selects an X-allele to inactivate, resulting in ratios of allelic-inactivation across embryos varying purely by chance (Fig. 3.1A). Closely following Mary Lyon’s discovery of XCI in 1961(Lyon 1961), it was recognized that the inherent embryonic stochasticity and permanence of XCI is the simplest explanation for the observed variability in XCI among adults and positions this adult variability as a window into embryonic events(Gandini et al. 1968; Gandini and Gartler 1969; Nesbitt 1971; Fialkow 1973; McMahon et al. 1983). With flipping coins as an example, it is much more probable to get 8 heads when flipping 10 coins than it is to get 80 heads when flipping 100 coins, i.e., the variability in heads-to-tails ratios is directly related to the number of coins flipped. In other words, the variability of XCI ratios in a population of female mammalian embryos is determined by the number of cells present at the time of XCI (Fig. 3.1A). Combined with the inheritance of allelic-inactivation through each cell’s lineage, quantifying XCI variability in adults can be taken as an approximation of embryonic XCI variability and used to infer cell counts at the time of XCI or other early lineage decisions(Nesbitt 1971; Werner et al. 2022) (Fig. 3.1D). Models of stochasticity have been used to infer cell counts during embryonic events in human and mice populations for decades(Sun et al. 2021; Gandini et al. 1968; Nesbitt 1971; McMahon et al. 1983; Werner et al. 2022; Bittel et al. 2008), with opportunities to assess the merits of stochastic models so far lacking in other mammalian species.

In addition to stochasticity, genetic effects can influence the choice of allelic inactivation and contribute to population variability in XCI ratios. The choice of allelic-inactivation during XCI is determined by the cis-acting long non-coding RNA (lncRNA) XIST(Brown et al. 1992), which coats its corresponding X-allele and initiates various transcriptionally silencing epigenetic modifications(Dossin and Heard 2021; Dixon-McDougall and Brown 2022). Heterozygous variants that impact XIST expression can bias the choice of allelic inactivation(Plenge et al. 1997). An extensively documented example of such an effect is the preferential inactivation of specific X-alleles in heterozygous laboratory mice(Cattanach and Isaacson 1965). Depending on the parental strains, inbred mice genomes exhibit a specific order of preferential allelic X-inactivation that is dependent on the X-chromosome controlling element (XCE) allele carried by each parent(Sun et al. 2021; Calaway et al. 2013). In humans, evidence for genetic influence on XCI is largely derived from small family studies and difficult to differentiate from potential disease effects, with no robust evidence supporting the broader allelic effects observed in mouse populations(Peeters et al. 2016; Bolduc et al. 2008). Another form of genetic influence on XCI is allelic selection, whereby natural genetic variation or disease-causing variants exert a selective effect on the X-alleles(Belmont 1996; Migeon 1998). Evidence supporting allelic selection in human populations is primarily limited to disease cases(Migeon 1971; Migeon et al. 1981; Devriendt et al. 1997; Plenge et al. 2002) or large genetic aberrations(Schmidt and Sart 1992), with broader evidence of allelic selection through natural variation remaining elusive. In general, the relative merits of genetic influences or models of stochasticity for explaining population XCI variability in mammals are difficult to resolve with available data currently limited to mouse and human populations.

While molecular studies of XCI enjoy cross-species comparisons enabled via embryonic models(Ramos-Ibeas et al. 2019; Magaraki et al. 2019; Yu et al. 2020; Okamoto et al. 2021), evaluations of XCI variability in adult populations across species is historically absent, largely due to prior technological limitations. Traditional methods for assessing XCI ratios typically rely on known polymorphic heterozygous X-linked regions(Amos-Landgraf et al. 2006; Allen et al. 1992) or personalized genetic information(Shvetsova et al. 2019; Szelinger et al. 2014), both methodological bottlenecks when studying population XCI variability at scale. Recently, we developed a method to estimate XCI ratios from bulk RNA-sequencing samples aligned to a reference genome(Werner et al. 2022) (Fig. 3.1B). Our approach leverages natural genetic variation to sample X-linked heterozygosity and eliminates the requirement for costly phased or strain specific genetic information. Importantly, our method is applicable to any bulk RNA-sequencing sample and reference genome, enabling us to utilize the vast amount of publicly available mammalian data (Fig. 3.1C), making cross-mammalian analysis of population XCI variability at scale feasible.

In this study, we source female annotated bulk RNA-sequencing samples across 8 non-human mammals from the Sequencing Read Archive (SRA), resulting in a total of 19,180 initial samples (Fig. 3.1C), including human samples from the GTEx(Lonsdale et al. 2013) dataset. We employ a rigorous sample processing pipeline for deriving high confidence calls of heterozygous SNPs from RNA-sequencing data controlling for reference bias and gene expression (Fig. 3.1C, see methods). We use X-linked sample heterozygosity to model the XCI ratio of individual samples (Fig. 3.1B) and investigate potential genetic correlates with XCI ratio variability across our mammalian populations. We start by establishing the population-level XCI ratio distributions for all nine mammalian species and use models of embryonic stochasticity to predict the number of cells fated for embryonic lineages (Fig. 3.1D, Fig. 3.2). We then investigate how broad genetic diversity, as indicated by measures of inbreeding (Fig. 3.3), as well as specific individual variants (Fig. 3.4), may impact population XCI variability. Overall, our analyses explore how both models of stochasticity and genetic factors can explain population XCI variability across 9 mammalian species.

3.4. Results

3.4.1. Reference aligned RNA-sequencing data enables scalable modeling of XCI ratios

X-linked allelic expression in a bulk RNA-sequencing (RNA-seq) sample is expected to reflect the XCI ratio of the sampled tissue (Fig. 3.1B). Explicitly, the sampled X-linked allelic reads are expected to follow a binomial distribution dependent on the number of sampled reads and the XCI ratio of the sample (see methods). We employ a maximum-likelihood approach to fit distributions to the observed allelic ratios of multiple heterozygous single nucleotide polymorphisms (SNPs, minimum of 10) per sample to compute estimates of XCI ratios, with special considerations when using reference-aligned data (Fig. 3.1B).

A reference genome includes both paternal and maternal SNPs for any given individual RNA-seq sample, leading to reference allelic expression ratios that represent allelic expression of both parental X-alleles (Fig. 3.1B). Folding the distribution of reference allelic-expression ratios around 0.50 aggregates data across both alleles and enables a robust estimate of the XCI ratio magnitude for the bulk RNA-seq sample (Fig. 3.1B). We fit folded-normal distributions to the reference allelic expression ratios, which serve as a continuous approximation of the underlying folded-binomial distribution. The mean of the fitted distribution is considered the estimate of the XCI ratio (Fig. 3.1B). We also incorporate specific steps to address confounding factors that can impact X-linked allelic expression, including reference bias and escape from XCI(Bonora and Disteche 2017; Fang et al. 2019) (Supp. Figs. 3.1-3.2, see methods). Of note, we find the strongest signals of escape from XCI near chromosomal ends across all species (Supp. Fig. 3.2), suggesting escape within pseudo-autosomal regions is conserved across mammals(Bonora and Disteche 2017; Posynick and Brown 2019). Previously, we validated our SNP filtering and XCI modeling approach using phased RNA-seq data (where haplotype information is known for each variant) from the EN-TEx consortium(Rozowsky et al. 2023), achieving nearly perfect agreement in XCI ratio estimates for samples with folded XCI ratios of 0.60 or higher, demonstrating the robustness of our approach.

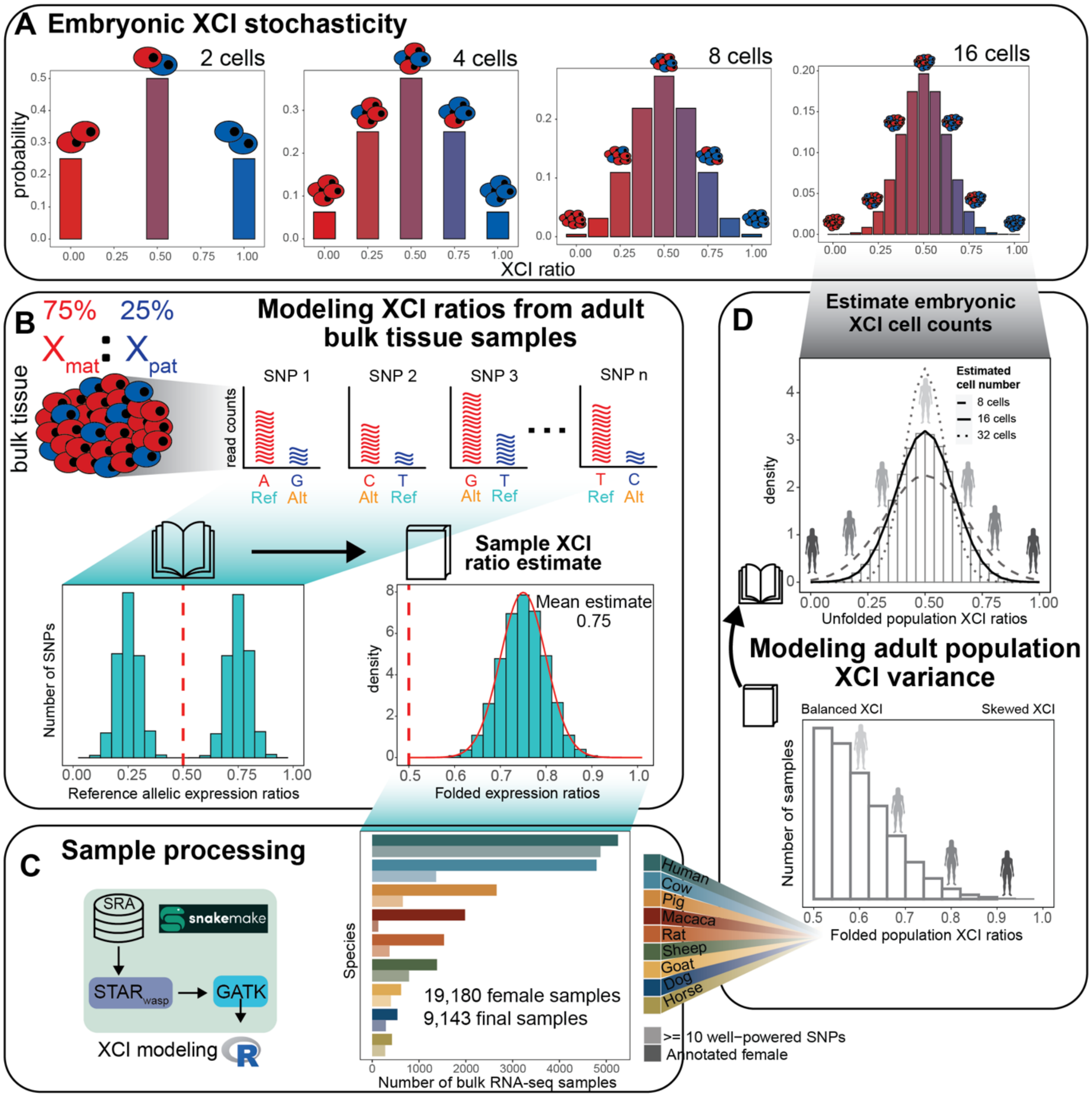


Figure 3.1: ***Reference aligned RNA-sequencing data enables scalable modeling of XCI ratios***
A Schematic demonstrating the relationship between the number of cells present at the time of XCI and the probability of all possible XCI ratios. Increased cell numbers result in decreased XCI ratio variance.
B Schematic for modeling XCI ratios from bulk reference-aligned RNA-seq data. The reference SNPs will contain both maternal and paternal SNPs, representing allelic expression from both parental haplotypes. Folded normal models are fit to the folded reference allelic expression ratios (like folding a book closed), with the mean of the maximum-likelihood distribution as the sample XCI ratio estimate.
C Schematic for sample processing (genome alignment and variant identification) and a bar graph depicting the number of annotated female samples initially downloaded for each species (bold color), with the number of samples per species with at least 10 well-powered SNPs for XCI ratio modeling after processing (faded color).

By calling SNPs from RNA-seq reads and employing folded distributions to model reference-aligned allelic expression, we can estimate the magnitude of XCI in any female mammalian bulk RNA-seq sample. We source female annotated bulk RNA-seq samples of 8 non-human mammalian species from the SRA database (Fig. 3.1C), additionally including cross-tissue human samples from the GTEx dataset. After processing, the number of samples with a minimum of 10 well-powered SNPs for estimating XCI ratios are 130 macaca (mean of 28 SNPs +- 17 SD), 275 horse (mean of 54 SNPs +- 36 SD), 269 dog (mean of 29 SNPs +- 13 SD), 328 rat (mean of 26 SNPs +- 13 SD), 383 goat (mean of 34 SNPs +- 14 SD), 624 pig (mean of 50 SNPs +- 28 SD), 731 sheep (mean of 79 SNPs +- 42 SD), 1328 cow (mean of 32 SNPs +- 19 SD), and 4877 human (mean of 56 SNPs +- 23 SD, 314 total individuals) samples (Fig. 3.1C, Supp. Fig. 3.1). Aggregating reference allelic expression ratios for samples with similar estimated XCI ratios (0.05 bins) clearly reveals the expected haplotype expression distributions, demonstrating the applicability of folded models (Supp. Fig. 3.3). Following XCI ratio modeling, we then generate population-level distributions by unfolding the distribution of folded XCI ratio sample estimates per species (Fig. 3.1D).

To ensure the allelic variability we report from X-linked SNPs is specific to XCI, we estimate autosomal allelic imbalances for all samples using the same pipeline and approach as for the X-chromosome analysis (Supp. Fig. 3.4, see methods). Comparing allelic imbalances across the two autosomes closest in size to the X-chromosome reveals the vast majority of samples across all species are biallelically balanced for autosomal expression, as expected (Supp. Fig. 3.4). Several species (Pig, Cow, Goat, Rat, Sheep, and Dog) exhibit small subsets of samples that are consistently imbalanced across the two autosomes and the X-chromosome, indicative of a global influence on allelic-expression independent of XCI (Supp. Fig. 3.4). When comparing the SNP allelic-expression ratios across samples that exhibit autosomal imbalances or not, we identify pervasive reference bias at the SNP-level as the cause for the global allelic imbalances (Supp. Fig. 3.4B). The samples with global allelic imbalances are excluded from all downstream analysis, ensuring the population distributions of XCI ratios reflect variability specific to XCI.

3.4.2. Models of embryonic stochasticity explain adult population XCI variability

After generating population distributions of XCI ratios for the 9 mammalian species, we next explore how well models of embryonic stochasticity explain the observed adult XCI ratio variability. The initial variability in XCI ratios among mammalian embryos is dependent on the number of cells present during XCI (Fig. 3.1A), where adult variability can be modeled to infer embryonic cell counts. An important consideration when estimating embryonic cell counts from adult XCI variability is the lineage specification of extra-embryonic and embryonic tissues, which may coincide with the timing of XCI and vary depending on the species. If XCI occurs after the lineage decision, the variability in XCI ratios within the embryonic lineage is determined by the number of cells fated for embryonic development at the time of XCI. On the other hand, if the order of events is reversed, XCI variability within the embryonic lineage is influenced by both the initial stochasticity of XCI and the stochasticity associated with cell sampling during the extra-embryonic/embryonic lineage decision. The ordering of these lineage events cannot be resolved without cross-tissue sampling of both the extra-embryonic and embryonic tissues. Therefore, estimating cell counts based solely on adult tissues provides an approximation of the number of cells fated for the embryonic lineage, representing the last common lineage decision for all the sampled adult cells.

Figure 3.2 A presents the unfolded population distributions of XCI ratios in the 9 mammalian species we sampled, ranging from the least variable (macaca) to most variable (dog). We fit normal distributions as continuous approximations to the underlying binomial

distribution that defines the relationship between cell counts and XCI ratio variability (Fig. 3.1A, see methods). We focus on the tails of the distributions, as our previous validation using phased data indicated increased uncertainty for folded XCI ratio estimates between 0.5-0.6, which translates to unfolded estimates between 0.4-0.6. At a broad level, population XCI ratio variability varies substantially across the sampled mammalian species. Our estimates for the number of cells fated for embryonic lineages include 65 (macaca), 31 (rat), 23 (pig), 16 (goat), 15 (horse), 14 (sheep), 14 (cow), 13 (human) and 8 (dog) cells, with associated 95% confidence intervals presented in figure 3.2B. The error between the empirical XCI ratio distributions and the normal fitted distributions is strikingly small, with a mean of 0.00538 (+- 0.0101 SD) across the species (Supp. Fig. 3.5). This demonstrates models of embryonic stochasticity can explain observed XCI ratio variability in adult populations exceptionally well.

For the least and most variable species (macaca and dog), the estimated autosomal imbalances offer additional context for the reported XCI population variability. The reported X-linked variability in macaca is in excess to the reported autosomal allelic variability (Supp. Fig. 3.4). This demonstrates the X-linked population variability for macaca, while strikingly small, is specific to XCI and informative for estimating cell counts. On the other hand, the dog population is the only one that contains samples with strong allelic imbalances on only one autosome, where autosomal imbalances in all other species are global (Supp. Fig. 3.4). This is suggestive of broader genomic incompatibilities within the dog population. The reported X-linked population variability in dog is likely a combination of XCI and broader allelic incompatibilities, positioning our estimate of 8 cells as a likely underestimate due to excess variability outside of XCI.

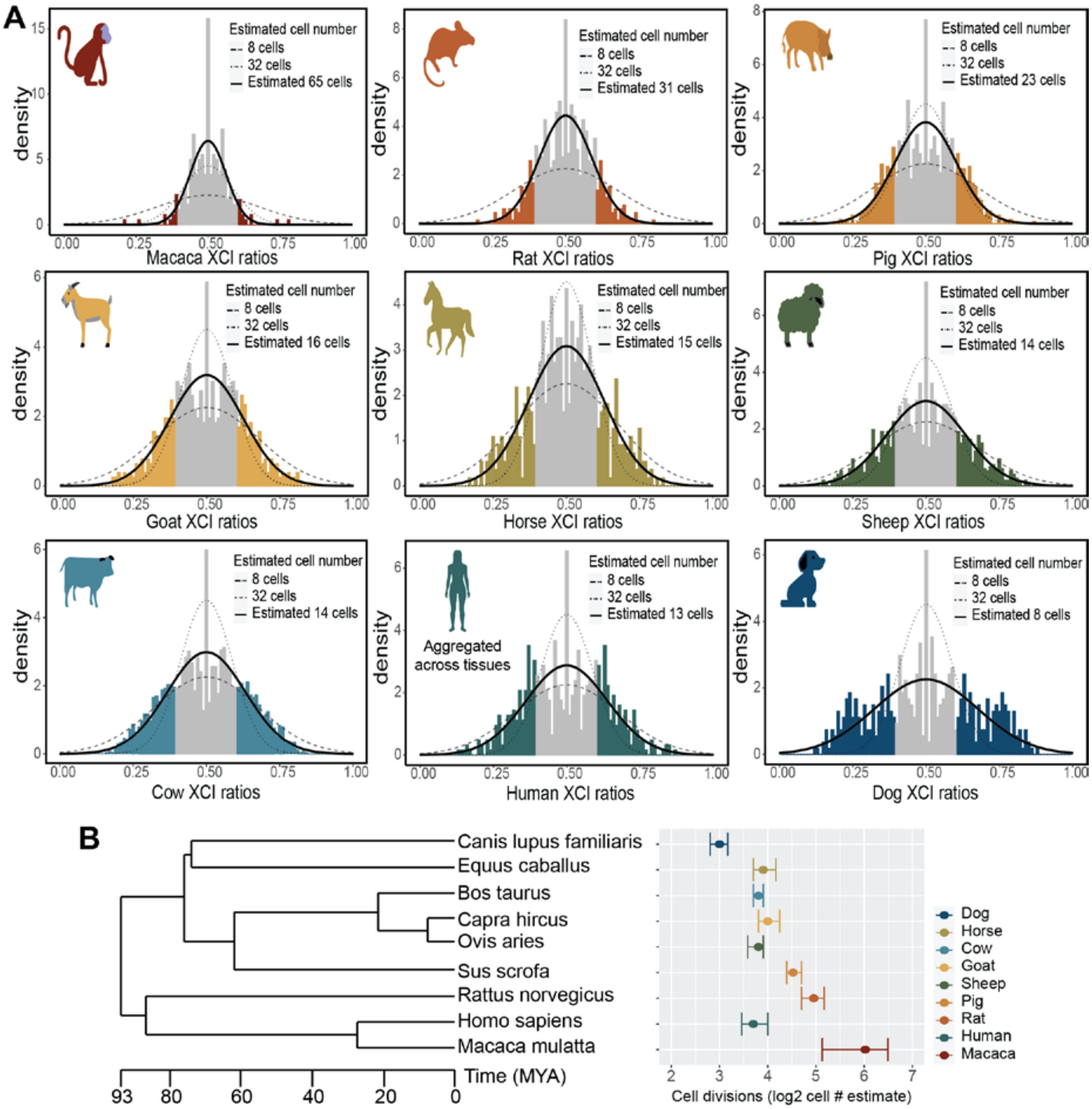


Figure 3.2: ***Models of embryonic stochasticity explain adult population XCI variability***
A Unfolded distributions of XCI ratios per species, with the maximum-likelihood normal distribution depicted in bold, fitted to the tails of the distributions (shaded in sections of the distributions).
B Phylogenetic tree of the sampled mammalian species with their estimated embryonic cell counts on a log-2 scale, depicting the number of cell divisions that separate the estimated cell counts between the species. Error bars are 95% confidence intervals around the cell number estimate.

Modeling XCI ratio variability across numerous species allows comparisons in light of evolution for determining generalizable or species-specific characteristics in XCI. Broadly, we demonstrate XCI ratios are variable in each species we assess, revealing variability in XCI ratios itself as a conserved characteristic of XCI. The exact variance in XCI ratios varies across the species, with differences in the timing of XCI and/or embryonic/extra-embryonic lineage specification (differences in cell counts) as one putative explanation. We compare our estimated cell counts to the evolutionary relationships among the species we assess (Fig. 3.2B), suggesting that variability in timing for these early embryonic events are recent evolutionary adaptations. This is highlighted by the large differences in cell counts between macaca and humans. When viewed through the lens of cell divisions (log2 of the estimated cell counts, Fig. 3.2B), the differences in XCI ratio variability among the species can be explained by differences in only 1 to 3 cell divisions, a narrow developmental window. This demonstrates even slight changes in the timing of XCI or embryonic/extra-embryonic lineage specification across mammalian species can produce large differences in population XCI ratio variability, as explained through the inherent stochasticity of XCI.

3.4.3. XCI ratios are not associated with X-linked heterozygosity

After determining stochastic models can explain population XCI ratio variability across mammalian species, we turn to testing whether we can identify any genetic correlates with XCI ratios. Our approach leveraging natural genetic variation to quantify XCI ratios enables the opportunity to assess an expansive catalog of genetic variants for associations with XCI ratios across mammalian species (10,735 macaca SNPs, 12,024 rat SNPs, 23,603 pig SNPs, 16,123 goat SNPs, 10,281 horse SNPs, 53,505 sheep SNPs, 18,509 cow SNPs, 16,168 human SNPs, and 10,050 dog SNPs). One putative genetic contribution to XCI ratio variability is allelic selection during development, where increased X-linked heterozygosity (i.e., genetic distance), is more likely to produce selective pressures between the two X-alleles. It follows that samples with higher X-linked heterozygosity would be expected to exhibit more extreme XCI ratios.

We score X-linked heterozygosity per sample as the ratio of the detected SNPs within a sample to the number of unique SNPs identified across all samples, relative for each species (Fig. 3.3A). This quantification also serves as a measure of inbreeding, with decreased heterozygosity associated with a higher degree of inbreeding(Miller et al. 2014). The trend in heterozygosity across species is as expected, with rats (likely laboratory strains) as the most inbred (Fig. 3.3A). Next, we examine the correlations between sample heterozygosity and the estimated XCI ratio, as well as the estimated variability in allelic expression per sample (mean and standard deviation of the fitted folded-normal distribution per sample, Fig. 3.3B). Across all species, X-linked heterozygosity showed a near-zero correlation with the estimated XCI ratio, indicating a lack of association between X-linked genetic variability and XCI ratio variability (Fig. 3.3B). However, we observe moderate correlations between sample heterozygosity and the estimated variability in allelic expression in three species: rat (corr: 0.576), macaca (corr: 0.459), and cow (corr: 0.364), notably the most inbred species (Fig. 3.3A, Supp. Fig. 3.6). The increased variability in allelic expression present only within the most inbred species could potentially reflect genomic incompatibility between parental haplotypes(Shorter et al. 2017) rather than a direct genetic effect on XCI.

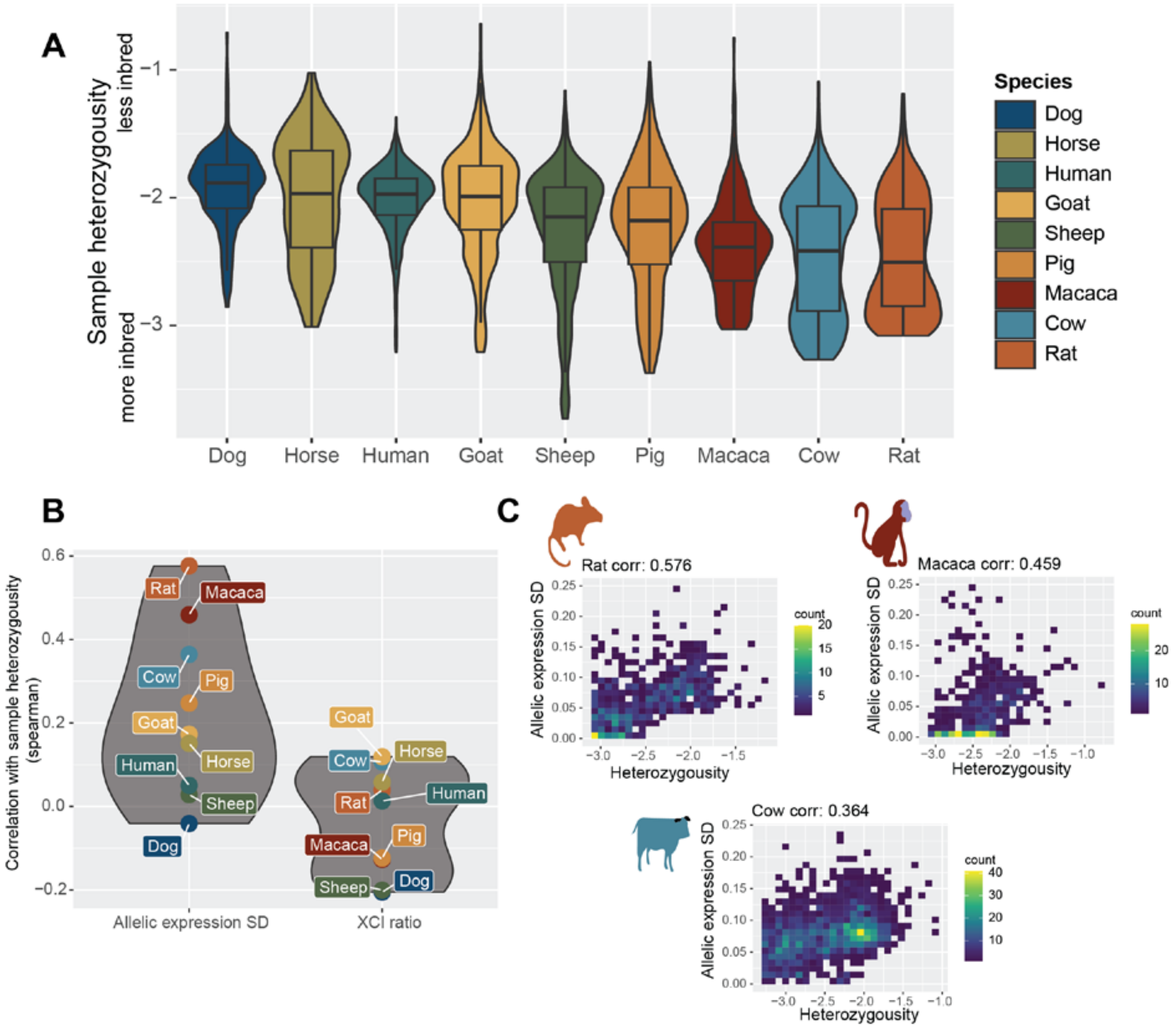


Figure 3.3: ***XCI ratios are not associated with X-linked heterozygosity***

A Distributions of sample X-linked heterozygosity per species ordered by the median value. The y-axis is in log-10 scale, depicting the ratio of SNPs per sample to all unique identified SNPs per species. Boxplots depict the distributions' quartiles.

B The spearman correlation coefficients between sample X-linked heterozygosity and either the estimated standard deviation (SD) in X-linked allelic expression or the estimated XCI ratio of the sample (the SD and mean of the maximum-likelihood folded-normal model per sample).

C 2D Scatter plots of sample heterozygosity compared to the sample estimated X-linked allelic expression SD for the three species with moderate correlation coefficients. Color bars represent the number of samples in each 2D bin. Plots for the other species are in Supp. Fig. 3.6.

3.4.4. Low frequency variants exhibit moderate associations with XCI ratios

After investigating relationships between genetic variation and XCI ratios at a broad level across the whole X-chromosome, we next asked if individual variants might be associated with extreme XCI ratios. Variants that affect the expression and/or function of the genetic elements that control XCI can result in highly skewed XCI ratios, as documented in human studies(Plenge et al. 1997). This can also occur in other X-linked genes, if the resulting differential in gene activity exerts a selective pressure across the X-alleles, as documented in disease cases(Migeon 1998; Belmont 1996). We test the association between XCI ratios and individual variants for all variants detected in each species with a minimum of 10 samples, quantified through the area-under-the-receiver-operating-curve statistic (AUROC). For each species, we rank the samples based on their estimated XCI ratio and score the placement of samples carrying a given variant within the ordered list (Fig. 3.4A). If all the samples with that variant are at the top of the ordered list, the XCI ratio can be said to have perfectly predicted the presence of that variant, quantified with an AUROC of exactly 1. An AUROC of 0.50 indicates the XCI ratio performs no better than random chance for predicting the presence of the variant.

The distribution of AUROCs for each species show striking similarities to a null comparison (Fig. 3.4B, see methods), indicating a pervasive lack of association between XCI ratios and individual variants. However, a small subset of variants in each species exhibits moderate associations (AUROCs >= 0.75). By comparing each variant's AUROC with its frequency in the species, we find that the variants with moderate associations occur at low frequencies within the sampled populations (Fig. 3.4C, Supp. Fig. 3.7). We investigate whether this relationship is simply due to a lack in power with bootstrap simulations, demonstrating moderate AUROCs (>= 0.75) are not produced purely through small sample sizes (Supp. Fig. 3.7). Figure 3.4D displays these variants along with their gene annotations for each species. Notably, several genes in humans with moderate AUROCs have prior evidence for associations with skewed XCI, namely MECP2(Knudsen et al. 2006), IDS(Kloska et al. 2011) (also identified in macaca), IRAK1(Morcillo et al. 2020), and FLNA(Robertson et al. 2003). This suggests our analysis is able to recover putative examples of selection impacting XCI ratios via disease-variants, though with small effect sizes and low frequencies in our sampled population. In general, we are unable to identify strong associations between genetic variation and XCI ratios across all 9 mammalian species, both along the whole X-chromosome and for individual variants.

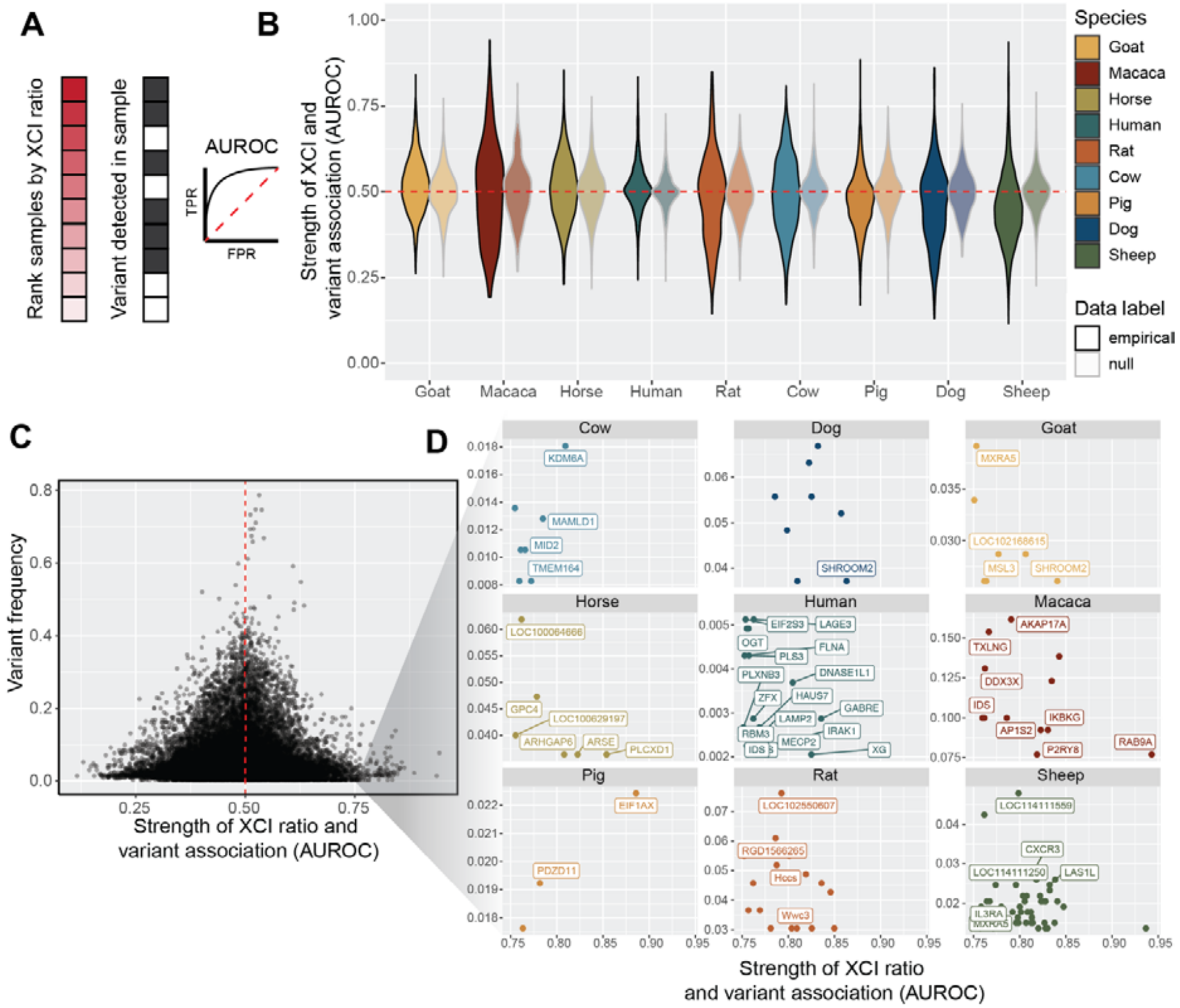


Figure 3.4 ***Low frequency variants exhibit moderate associations with XCI ratios***
A Schematic depicting the AUROC quantification for testing the association between individual variants and extreme XCI ratios. Samples are ranked by their estimated XCI ratio, with the dark shaded red squares representing samples with more extreme XCI ratios. The position of samples with a given individual variant (grey squares) within the ranked list is used to compute the AUROC statistic. A variant with an AUROC value of 1 means all samples with that variant were at the top of the ranked list, whereas an AUROC value of 0.5 represents a random ordering of samples within the ranked list.
B Distributions of variant AUROCs for each species compared to a species-specific null distribution of AUROC values (faded distributions, see methods), ordered by the mean value of the empirical distributions. The red dotted line depicts an AUROC of 0.50, performance due to random chance.
C Scatter plot of variant AUROCs compared to each variant's prevalence (percent of samples with that variant, relative for each species) for all variants across all species. The red dotted line depicts an AUROC of 0.50, performance due to random chance. A threshold of AUROC >= 0.75 was used to identify SNPs with moderate associations with XCI ratios.
D Scatter plots depicting the same information as in C for the variants with moderate associations with XCI ratios, but split by each species and including gene annotations. SNPs not within annotated genes are unlabeled. Gene labels not present due to overlapping labels are Macaca: ZBED1, Sheep: LOC101108113, LOC101115509, LOC101117055, LOC105605313, LOC121818231, PPP2R3B, PRKX)

3.5. Methods

3.5.1. Snakemake pipeline for RNA-seq alignment and variant identification

All non-human mammalian fastq data was downloaded from the Sequencing Read Archive (SRA, https://www.ncbi.nlm.nih.gov/sra), where only samples annotated as female were selected, using the metadata provided through SRA. Details for download and processing of the GTEx(Lonsdale et al. 2013) data can be found here(Werner et al. 2022). The entire sample processing pipeline uses a standard collection of bioinformatics software tools, all available for installation via Conda (STAR(Dobin et al. 2013) v2.7.9a, GATK(McKenna et al. 2010) v4.2.2.0, samtools(Li et al. 2009) v1.13, igvtools(Robinson et al. 2011) v2.5.3, and sra-tools 2.11.0). All Snakemake workflow rules, environment setup procedure, analysis commands and options, and underlying libraries are available on Github at https://github.com/gillislab/cross_mammal_xci , and https://github.com/gillislab/xskew. Briefly, a .fastq file acts as input, for either single- or pair-end sequencing experiments, and a .vcf and .wig file are produced as outputs for subsequent compiling of allele-specific read counts in R v4.3.0. The R script used for combining the .vcf and .wig information is also made available at https://github.com/gillislab/cross_mammal_xci/tree/main/R. Genome generation and alignment was performed with STAR, with the addition of the WASP(van de Geijn et al. 2015) algorithm for identifying and excluding reference biased reads. We extract chromosome-specific alignments from the .bam file (X chromosome or specific autosomes) and use GATK tools to identify heterozygous SNPs from that chromosome. The suite of GATK tools for identifying heterozygous variants from RNA-sequencing data was used following the GATK Best Practices recommendations. Specifically, the tools utilized include AddOrReplaceReadGroups -> MarkDuplicates -> SplitNCigarReads -> HaplotypeCaller -> SelectVariants -> VariantFiltration.

Reference genomes and gene annotations (.gtf files) for each species were sourced from the NCBI Refseq database (https://www.ncbi.nlm.nih.gov/refseq/). In each case the latest assembly version path was used, and the genomic.fna and genomic.gtf was downloaded. Annotated and indexed genomes were generated with STAR using --runMode genomeGenerate with default parameters.

3.5.2. SNP filtering

Only SNPs with exactly two identified genotypes were included for analysis and indels were excluded. We required each SNP to have a minimum of 10 reads mapped to both alleles for a minimum read depth of 20 reads per SNP. Gene annotations for all SNPs were extracted from the species-specific .gtf files. For XCI ratio modeling, we only used SNPs found within annotated genes. For any sample with multiple SNPs identified in a gene, we took the SNP with the highest read count to be the max-powered representative of that gene, so each individual SNP is representative of a single gene. In addition to implementing the WASP algorithm for excluding reference biased reads, we filter out SNPs within each species whose mean expression ratios across samples deviate strongly from 0.50 (mean allelic ratio < 0.40 and > 0.60, Supp. Fig. 3.1). This SNP filtering also excludes potential eQTL effects that may impact allelic-expression outside of the underlying XCI ratio.

3.5.3. Identifying and excluding chromosomal regions that escape XCI

We reasoned robust escape from XCI would produce more balanced biallelic expression in samples with skewed XCI. We performed an initial pass at XCI ratio modeling including all well-powered SNPs in a sample to identify samples with skewed XCI ratios (XCI ratios >= 0.70 for all species except rat and macaca, where a threshold of 0.60 was used due to a reduced incidence of skewed XCI in these species). Using the subset of skewed samples for each species, we averaged the folded allelic-expression ratios for all SNPs present in 1 mega-base (MB) bins across the X-chromosome (Supp. Fig. 3.2). Chromosomal-bins that displayed balanced allelic expression in opposition to the clearly skewed allelic expression of the rest of the chromosome were excluded from analysis. Specifically, chromosomal bins with an average allelic-expression < 0.65 for pig, goat, horse, sheep, and cow, < 0.60 in rat and macaca, and <0.675 in dog were excluded (Supp. Fig. 3.2) The ends of the X-chromosome in all species, except rat, demonstrated strong balanced biallelic expression, indicative of escape within putative pseudo-autosomal regions. We excluded any bin within these putative pseudo-autosomal regions regardless of average allelic expression. The escape threshold for dog was increased to exclude all bins within the dog putative pseudo-autosomal region.

3.5.4. Modeling XCI ratios with the folded-normal distribution

Starting with a single parental allele, the sampled maternal allelic-expression of a heterozygous X-linked SNP can be modeled with a binomial distribution, dependent on the ratio of active maternal X-alleles in the sample and the read depth of the SNP.

$$\frac{X_{mat}}{n_{reads}} \sim \frac{Bin(n_{reads}, p_{mat})}{n_{reads}};\ E\left[\frac{X_{mat}}{n_{reads}}\right] = p_{mat};\ \ Var\left(\frac{X_{mat}}{n_{reads}}\right) = \frac{p_{mat}(1-p_{mat})}{n_{reads}},$$

where X_{mat} is the number of maternal allelic reads, n_{reads} is the read depth of the SNP, and p_{mat} is the ratio of active maternal X-alleles. When aligned to a reference genome, the parental phasing information is lost and the allelic-expression of X-linked SNPs can instead be modeled with the folded-binomial model(Urbakh 1967; Gart 1970). We utilize the folded-normal model as a continuous approximation of the underlying folded-binomial distribution due to various factors not accounted for in the binomial model (varying SNP expression levels and potential eQTL effects). The probability of allelic-expression under the folded-normal model is defined as:

$$\Pr(x_{ratio}; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x_{ratio} - \mu)^2}{2\sigma^2}} + \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x_{ratio}+\mu-1)^2}{2\sigma^2}}, \text{for } \mu \in\ [0.50, 1]\ ,$$

where x_{ratio} is the folded allelic-expression ratio of a SNP, μ is the folded XCI ratio of the sample, and σ is the standard deviation of the folded-normal distribution. We utilize a maximum-likelihood approach (negative log-likelihood minimization of eq. 2) to fit folded-normal distributions to the observed folded allelic-expression ratios of at least 10 filtered SNPs per sample, taking the μ parameter of the maximum-likelihood folded-normal distribution as the folded XCI ratio estimate of the sample.

3.5.5. Modeling autosomal imbalances

The folded-normal model can also be applied to autosomal data to estimate allelic-imbalances. For each species, we extract chromosome-specific alignments from the .bam file for the two autosomes closest in size to the X-chromosome (Supp. Fig. 3.4). We employ the exact same processing pipeline and thresholds as used for the X-chromosome. Any sample that displayed an autosomal imbalance greater than or equal to a folded estimate of 0.60 (dotted lines in Supp. Fig. 3.4) on either autosome was excluded from downstream analysis.

3.5.6. Modeling population XCI variability with models of embryonic stochasticity

XCI is a binomial sampling event, where the number of cells choosing to inactivate the same X-allele follows a binomial distribution defined as:

$$X \sim Bin(n_{cells}, p_{inact}),$$

where X is the number of cells inactivating the same X-allele, n_{cells} is the number of cells present at the time of XCI, and p_{inact} is the probability of inactivation (0.50).

Embryonic XCI ratios can be modeled as:

$$\frac{X}{n_{cells}} \sim \frac{Bin(n_{cells}, p_{inact})}{n_{cells}}$$

We estimate n_{cells} by fitting normal distributions to the unfolded population XCI ratio distributions of each species, as a continuous approximation for the underlying binomial distribution. The variance of the normal distribution is defined as:

$$var_{normal} = Var\left(\frac{Bin(n_{cells}, p_{inact})}{n_{cells}}\right) = \frac{p_{inact}(1 - p_{inact})}{n_{cells}} = \frac{.5(1 - .5)}{n_{cells}}$$

We model population XCI ratios as:

$$\frac{X}{n_{cells}} \sim Norm(\mu, \sqrt{var_{normal}}\,),$$

where μ = p_{inact} = 0.50 and var_{normal} is computed for $n_{cells} \in [2, 200]$.

We identify the normal distribution with minimal sum-squared error between its CDF and the empirical population XCI ratio CDF, minimizing error over the tails of the distributions with percentiles <= 0.40 or >= 0.60 (Supp. Fig. 3.5). We compute 95% confidence intervals about the cell number estimate n_{cells} through bootstrap simulations. We sample with replacement from the empirical population XCI ratio distribution, matching the sample size of the original empirical population distribution, and fit a normal model to derive a bootstrap estimate of n_{cells}. We repeat this for 2000 simulations to generate a bootstrapped distribution of n_{cells}, from which we derive the 95% confidence intervals, defined as the interval where 2.5% of the bootstrapped distribution lies outside either end.

3.5.7. Measuring sample X-linked heterozygosity

We compute sample heterozygosity as the ratio of SNPs detected in a sample (20 read minimum) to the total number of unique SNPs identified across all samples for a given species. We quantify associations between X-linked heterozygosity and XCI ratios as the spearman correlation coefficient between the sample X-linked heterozygosity ratio and the fitted mean and variance of the maximum-likelihood folded-normal distribution of the sample (Fig. 3.3B-C, Supp. Fig. 3.6). We only consider samples with at least 10 detected SNPs.

3.5.8. Quantifying variant associations with extreme XCI ratios

We quantify the strength of XCI ratios as a predictor for the presence of a given variant through the AUROC metric. Given a ranked list of data (XCI ratios) and an indicator of true positives (samples with a given variant), the AUROC quantifies the probability a true positive is ranked above a true negative. An AUROC of 1 indicates all true positive samples were ranked above all true negative samples, demonstrating XCI ratios were a perfect predictor for the presence of that variant. An AUROC of 0.50 indicates random placement of true positives and negatives in the ranked list, demonstrating XCI ratios performed no better than random chance for predicting the presence of that variant. We compute the AUROC through the Mann-Whitney U-test, defined as:

$$AUROC = \frac{U}{n_{pos} + n_{neg}},$$

where U is the Mann-Whitney U-test test statistic, computed in R with wilcox.test(alternative = ‘two.sided’), n_{pos} is the number of true positive samples and n_{neg} is the number of true negative samples. We generate a null AUROC per variant by randomly shuffling the true positive and negative labels. The variant frequency is defined as the number of samples that carry a given variant over the total number of samples for a given species. The p-value for a given AUROC is the p-value associated with the Mann-Whitney U-test test statistic (U), where we determine significance as an FDR-corrected p-value <= 0.05. We perform FDR correction for all p-values computed for all variants across the 9 species through the Benjamini-Hochberg method, implemented in R via p.adjust(method = ‘BH’).

We estimate the power of each variant through bootstrap simulations. We randomly sample with replacement the XCI ratios of the true positive and true negative samples, those that either carry or do not carry a given variant. We match the sample size of the original true positive and negative labels. We compute a bootstrapped AUROC and p-value from the simulated data, repeating for 2000 simulations to compute a bootstrapped distribution of AUROCs. The AUROC power (Supp. Fig. 3.7B) is defined as the fraction of bootstrapped AUROCs that are significant, using a significance threshold of p-value <= 0.05. The AUROC effect size power (Supp. Fig. 3.7C) is defined as the fraction of bootstrapped AUROCs that are >= 0.75. We also report the variance of the bootstrapped AUROC distribution per variant in Supp. Fig. 3.7D. We exclude all variants classified as reference biased from Supp. Fig. 3.1, with the distributions of AUROCs for the reference biased and non-reference biased SNPs presented in Supp. Fig. 3.7E.

3.5.9. Software

All analysis was performed in R(R Core Team 2023) v4.3.0. All plots were generated using ggplot2(Wickham 2016) v3.4.2 functions. The phylogenetic tree in Fig. 3.2B was generated from TimeTree http://www.timetree.org/.

3.5.10. Data and Code availability

All associated code can be found at https://github.com/gillislab/cross_mammal_xci. This includes the snakemake pipeline used for processing the non-human mammalian data as well as all R notebooks used for data analysis and figure generation.

3.6. Supplemental figures

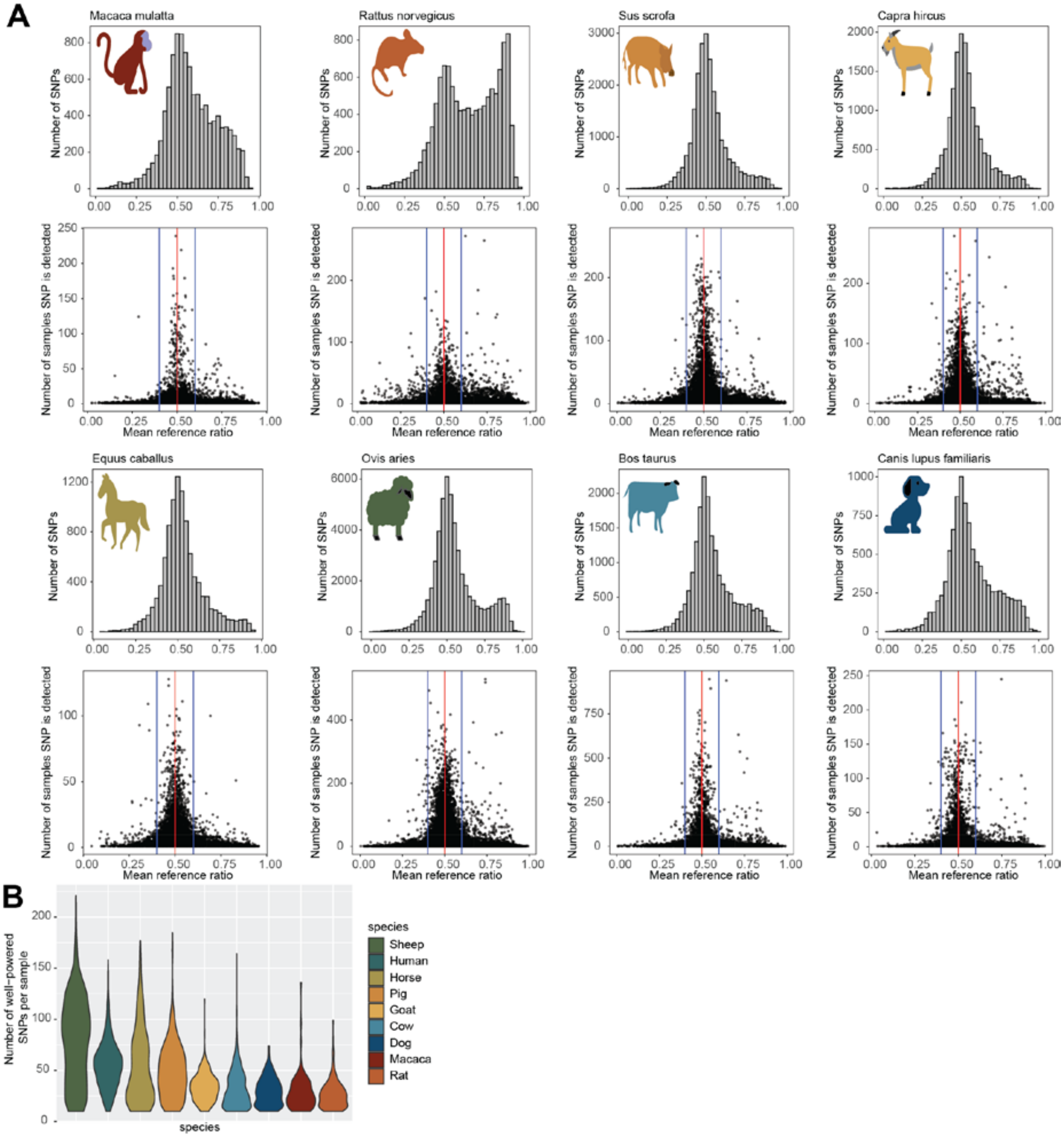


Figure S3.1: ***Reference bias varies across individual SNPs***

A Top histogram depicts the distribution of mean reference ratios for all detected SNPs in each species. The bottom scatter plot depicts the mean reference ratio against the sample size for each SNP. We exclude all SNPs from XCI ratio modeling whose mean reference ratio is < 0.40 or > 0.60 (blue lines), indicating consistent bias in allelic expression for either the alternate or reference allele.

B Violin plots depicting the distribution of the number of filtered SNPs (see methods) per sample for each species, where we require a minimum of 10 SNPs for XCI ratio modeling.

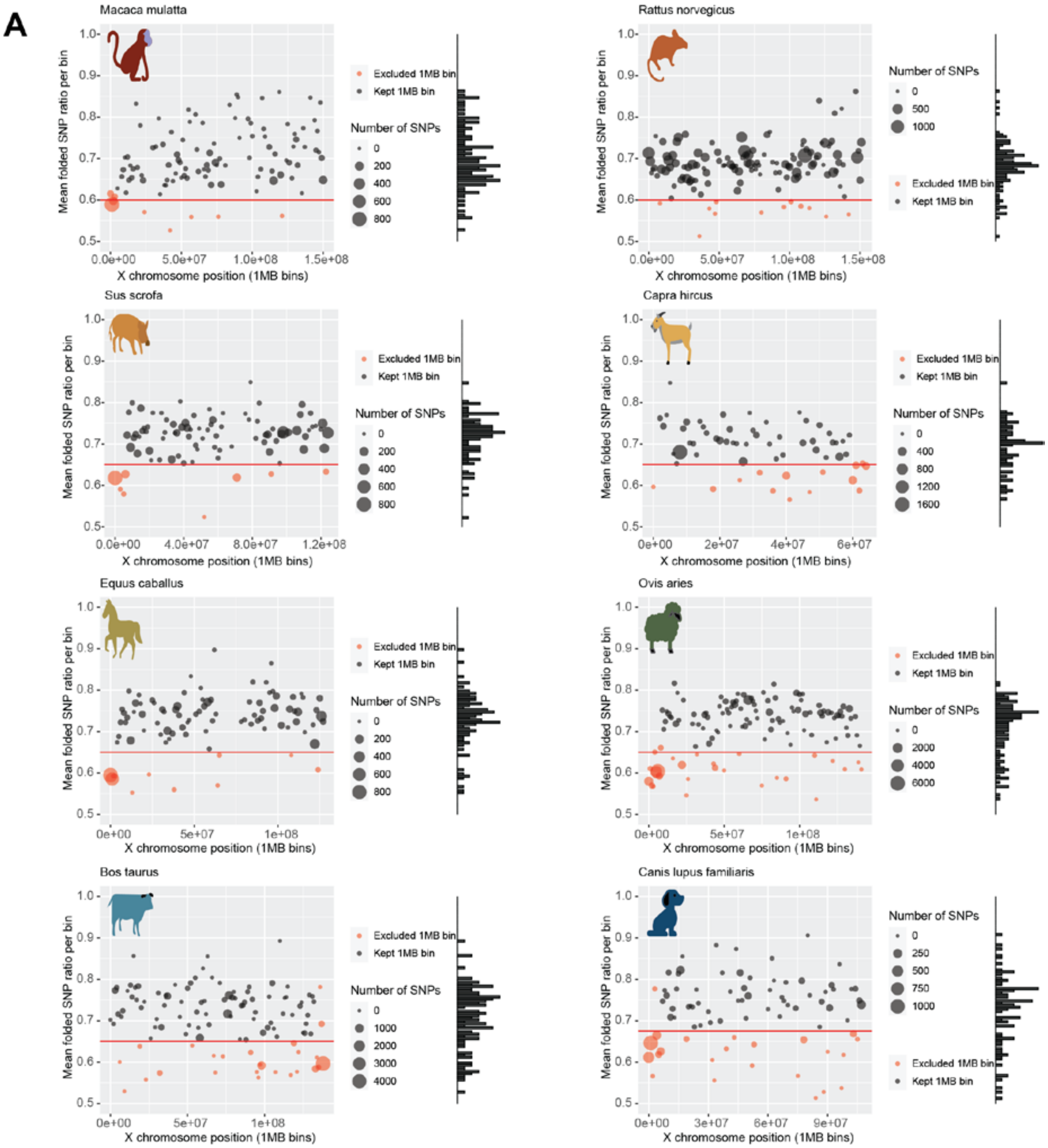


Figure S3.2: ***Escape from XCI is enriched in chromosomal ends***
A Scatter plots comparing the chromosomal location (1 mega-base bins) and the mean folded allelic expression ratio for all SNPs within each 1MB bin, derived from samples with skewed XCI (see methods). The marginal histogram depicts the distribution of mean folded allelic expression ratios per 1MB bin. The size of the data points corresponds to the number of SNPs in each 1MB bin. Red data points indicate 1MB bins that were excluded from analysis as probable escape regions, due to balanced allelic expression in samples with skewed XCI ratios. The chromosomal ends of all species, except rat, exhibit large clusters of SNPs with escape signal, likely pseudo-autosomal regions. The red lines depict the threshold of allelic-expression used to classify 1MB bins as escape or not.

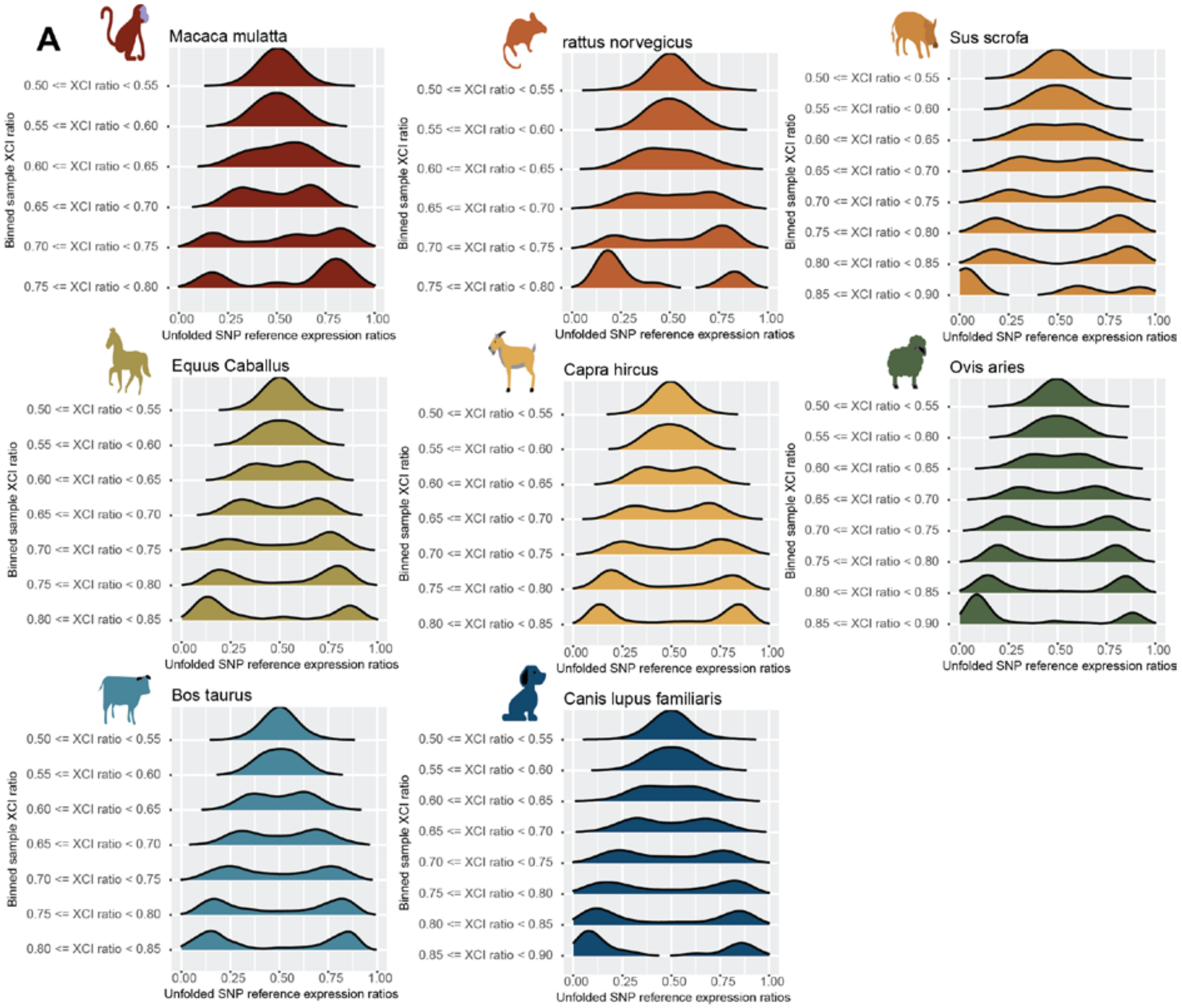


Figure S3.3: ***Reference allelic expression distributions exhibit bi-parental haplotype expression signatures expected of the X-chromosome***
A Density distributions of reference allelic expression ratios aggregated across samples binned by their estimated XCI ratio, ordered from balanced to more extreme XCI ratios (top to bottom). For samples with balanced XCI, the parental haplotypes cannot be distinguished, with clear separation of the parental haplotypes as the XCI ratio increases.

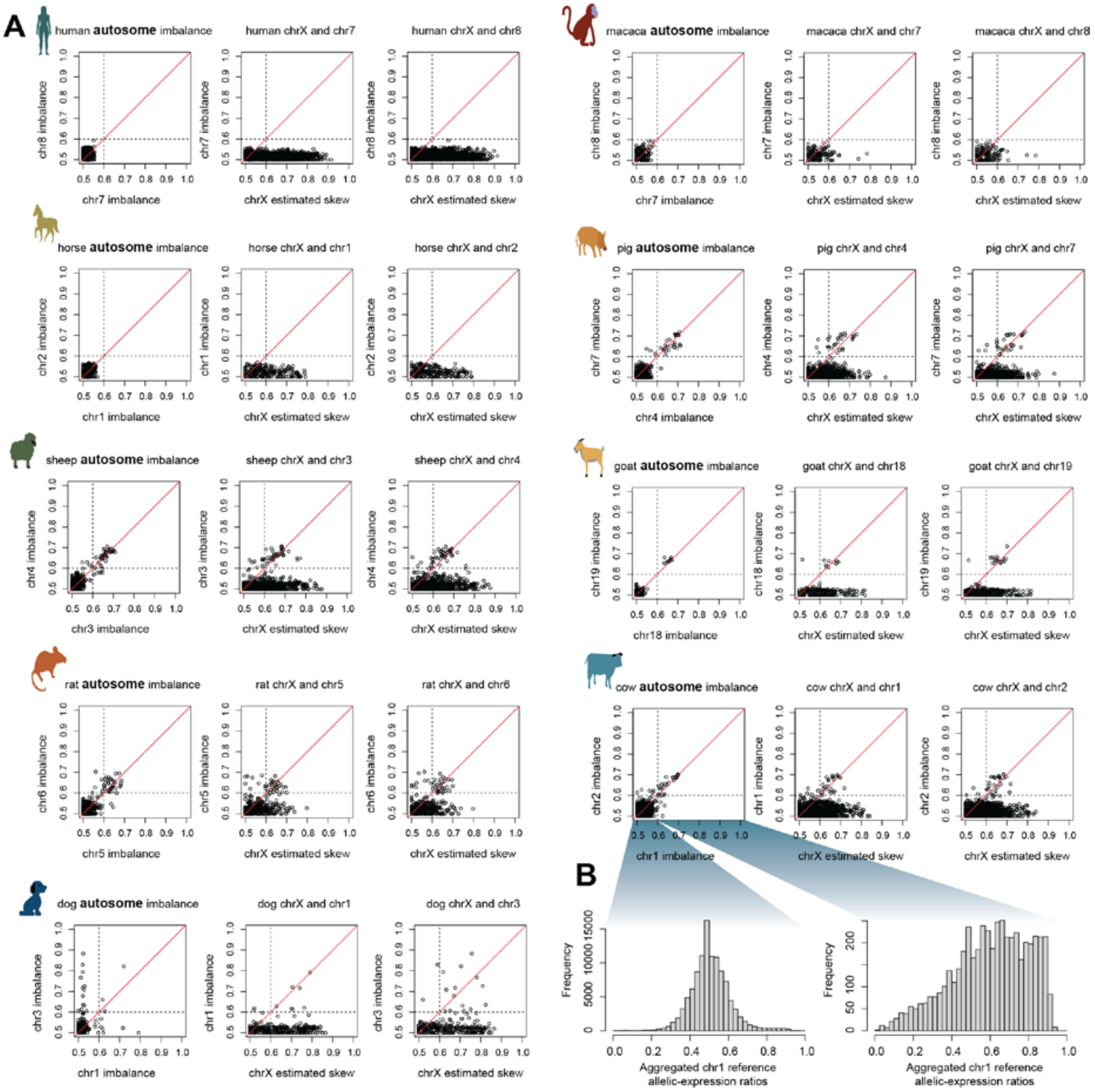


Figure S3.4: ***Comparing autosomal and X chromosome allelic imbalances***
A Each data point is a single individual sample. For each species, the first scatter plot compares the aggregated allelic imbalance of two autosomes (see methods). The following two scatter plots compare the aggregated allelic imbalance of an autosome and the X chromosome. Samples that exhibited imbalanced allelic expression on an autosome were excluded from analysis, using a threshold of an imbalance >= 0.60 (dotted lines).
B Histogram of the reference allelic-expression ratios of all chromosome 1 SNPs from the Cow samples with a chromosome 1 autosomal imbalance either < 0.60 (left) or >= 0.60 (right). The large autosomal imbalances can be attributed to extensive reference bias in allele-specific expression ratios. Cow results are representative of all other species.

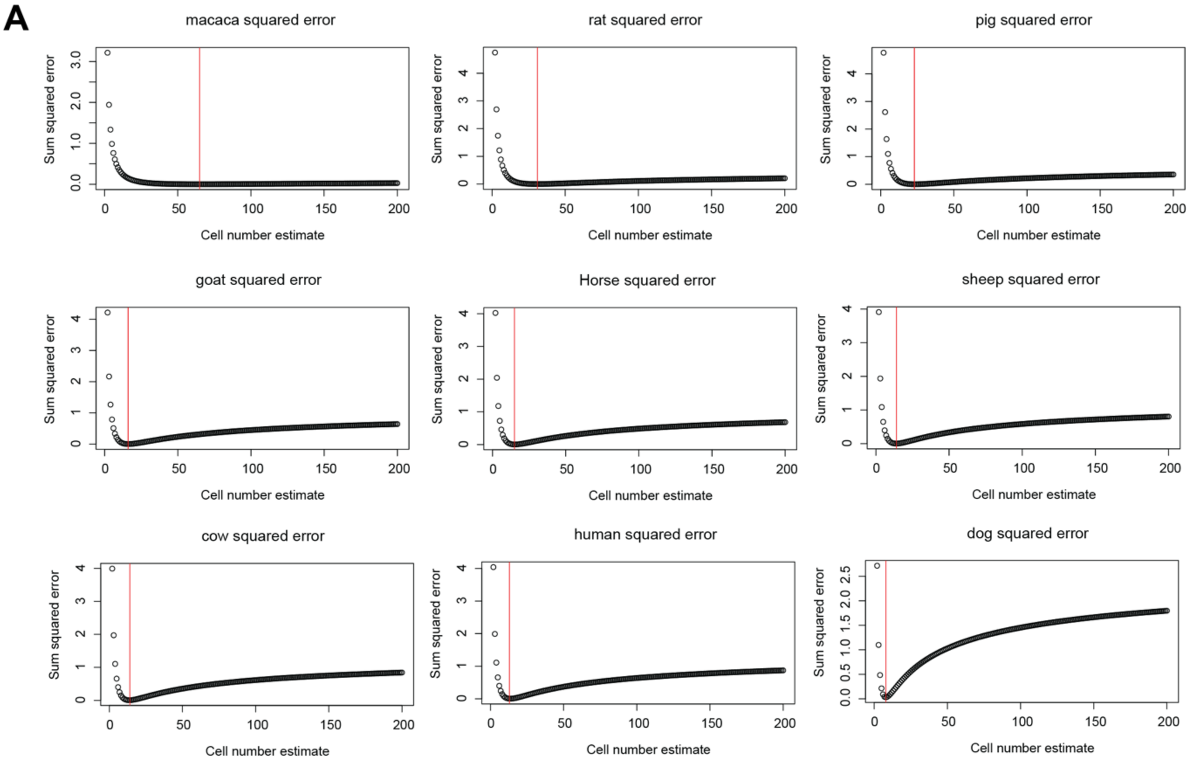


Figure S3.5: ***Estimating embryonic cell counts from population XCI ratio variance***
A Plots comparing the normal model associated with the estimated number of cells present during embryonic lineage specification (x-axis, see methods) to the sum of squared error between the percentiles of the tails of the empirical population XCI ratio distribution and the theoretical normal model (y-axis, see methods). The red lines depict the normal model with minimum error and the associated cell number estimate for each species.

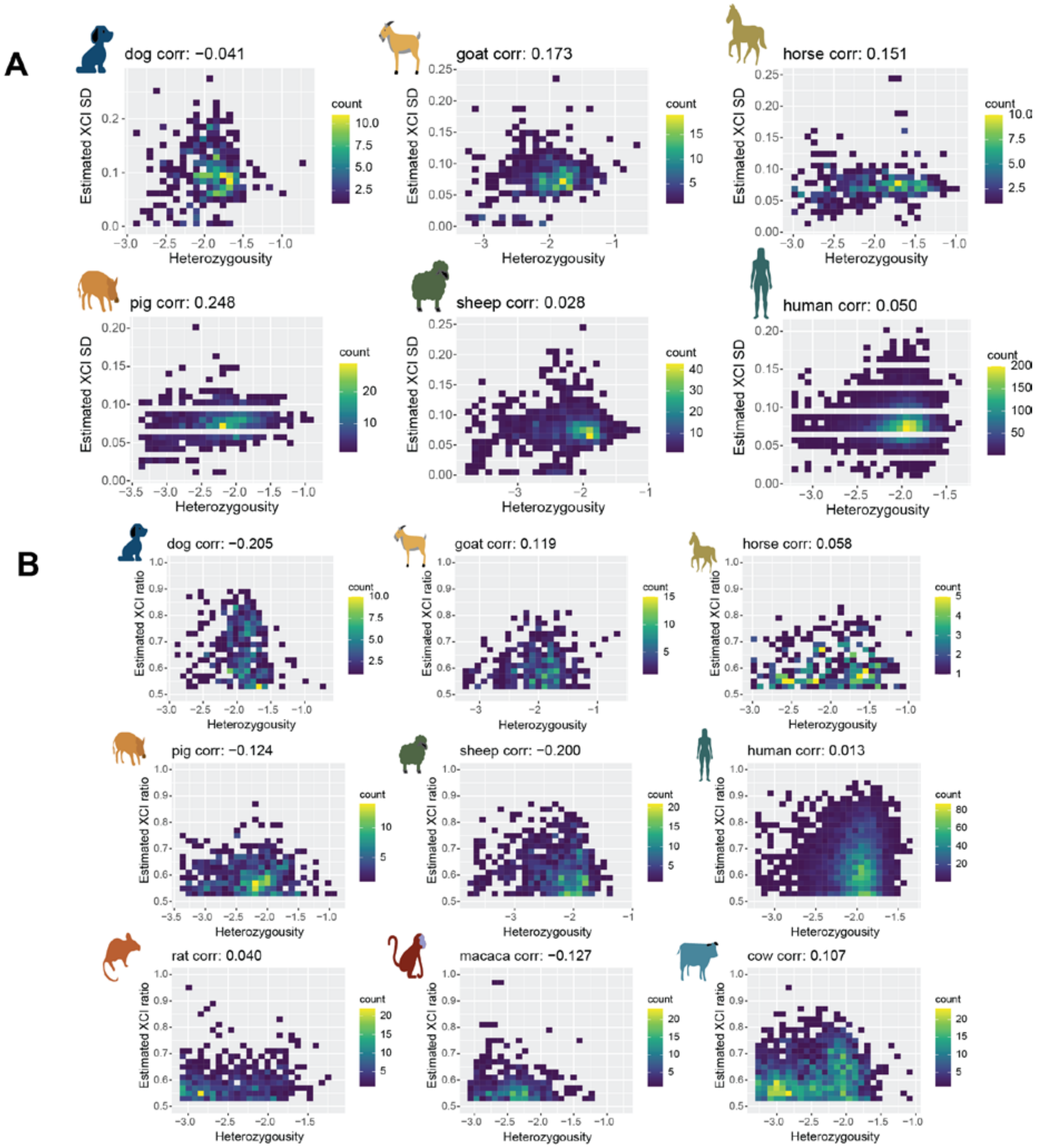

A
dog corr: −0.041
goat corr: 0.173
horse corr: 0.151
pig corr: 0.248
sheep corr: 0.028
human corr: 0.050
Estimated XCI SD
Heterozygousity
count
B
dog corr: −0.205
goat corr: 0.119
horse corr: 0.058
pig corr: −0.124
sheep corr: −0.200
human corr: 0.013
rat corr: 0.040
macaca corr: −0.127
cow corr: 0.107
Estimated XCI ratio
Heterozygousity
count

Figure S3.6: ***Species with no association between sample heterozygosity and variance in X-linked allelic expression***

A Binned scatter plots comparing the sample heterozygosity (log-10 of the number of SNPs per sample divided by the number of unique SNPs detected per species) to the estimated standard deviation (SD) in X-linked allelic expression (SD of the maximum-likelihood folded-normal distribution per sample). Spearman correlation coefficients are presented next to the species' names. Color bars represent the number of datapoints per 2D bin.

B Binned scatter plots comparing the sample heterozygosity (log-10 of the number of SNPs per sample divided by the number of unique SNPs detected per species) to the estimated XCI ratio (mean of the maximum-likelihood folded-normal distribution per sample). Spearman correlation coefficients are presented next to the species' names. Color bars represent the number of datapoints per 2D bin.

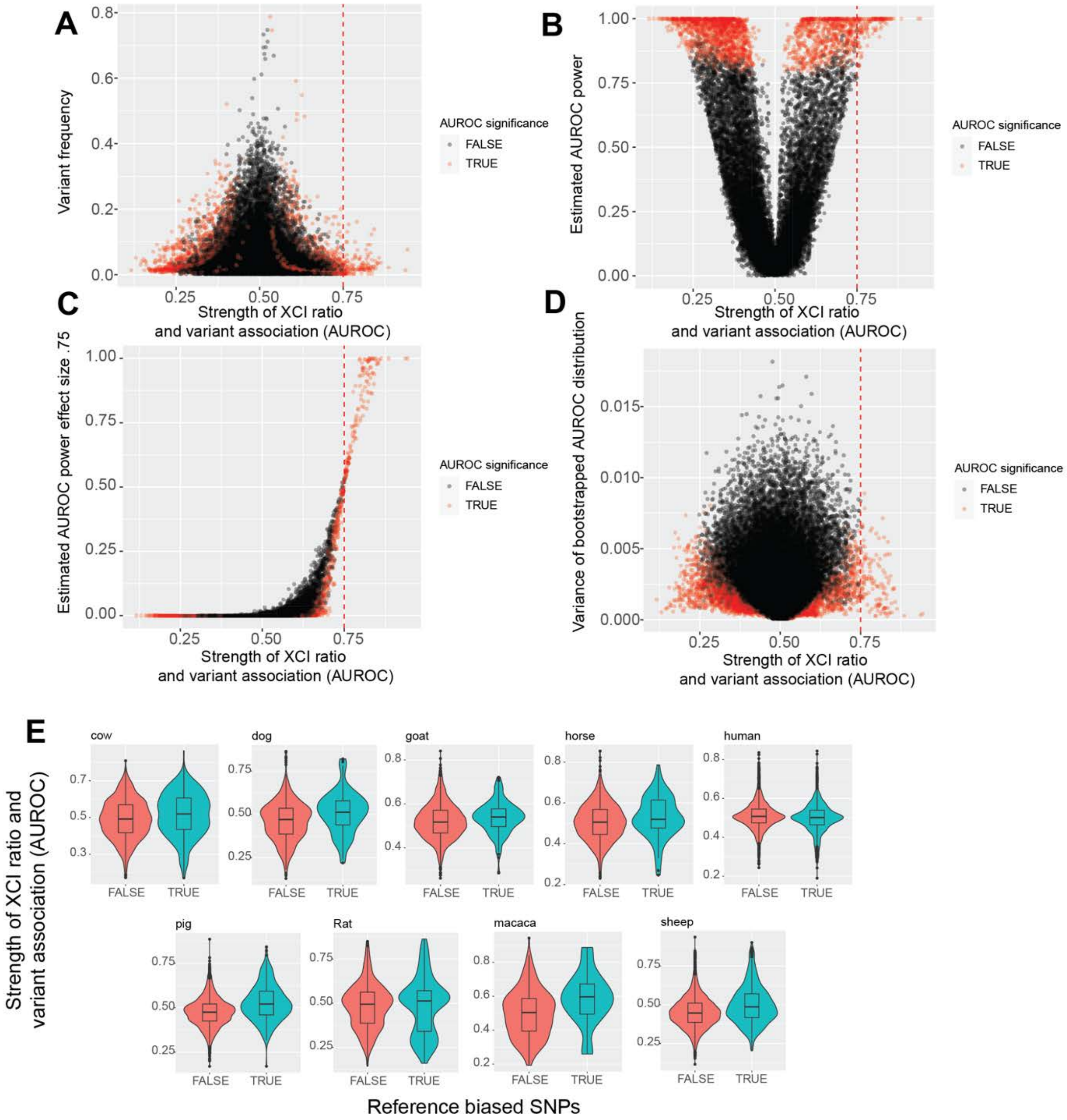

A
Variant frequency
0.8
0.6
0.4
0.2
0.0
0.25
0.50
0.75
Strength of XCI ratio and variant association (AUROC)
AUROC significance
FALSE
TRUE
B
Estimated AUROC power
1.00
0.75
0.50
0.25
0.00
0.25
0.50
0.75
Strength of XCI ratio and variant association (AUROC)
AUROC significance
FALSE
TRUE
C
Estimated AUROC power effect size .75
1.00
0.75
0.50
0.25
0.00
0.25
0.50
0.75
Strength of XCI ratio and variant association (AUROC)
AUROC significance
FALSE
TRUE
D
Variance of bootstrapped AUROC distribution
0.015
0.010
0.005
0.000
0.25
0.50
0.75
Strength of XCI ratio and variant association (AUROC)
AUROC significance
FALSE
TRUE
E
Strength of XCI ratio and variant association (AUROC)
cow
0.7
0.5
0.3
dog
0.75
0.50
0.25
goat
0.8
0.6
0.4
horse
0.8
0.6
0.4
0.2
human
0.8
0.6
0.4
0.2
pig
0.75
0.50
0.25
Rat
0.75
0.50
0.25
macaca
0.8
0.6
0.4
0.2
sheep
0.75
0.50
0.25
FALSE
TRUE
Reference biased SNPs

Figure S3.7: ***Low frequency variants are powered to detect significant associations with XCI ratios***
A Scatter plot comparing the initial AUROC against variant frequency for all variants across all species. Statistical significance of AUROCs is determined by an FDR-corrected p-value <= 0.05. The red dotted line in all 4 figure panels represents the AUROC threshold used to determine individual variants with a moderate association with XCI ratios.
B Scatter plot comparing the initial AUROC against the estimated power to detect a significant effect for each variant. Power was estimated through bootstrap simulations using a significance threshold of p-value <= 0.05, see methods.
C Scatter plot comparing the initial AUROC against the estimated power to detect an AUROC with effect size 0.75 or greater. Power was estimated through bootstrap simulations, see methods.
D Scatter plot comparing the initial AUROC against the variance of the bootstrapped distribution of AUROCs for each variant, see methods.
E Violin and boxplots depicting the distributions of AUROCs for the SNPs classified as either reference biased or not from the analysis in Supp. Fig. 1. Boxplots depict the quartiles of the distributions.

3.7. Chapter 3 summary:

In this work, we compile population distributions of XCI ratios across 9 mammalian species and assess both stochastic and genetic factors for explaining the observed population variability in XCI ratios. We reveal that mammalian populations generally vary in XCI ratios, with species ranging from very low variability (Macaca) to very high variability (Dog). We demonstrate that models of embryonic XCI stochasticity explain population XCI ratio variability exceptionally well across all species and provide estimates for the number of cells that must have been present in the embryonic epiblast during XCI for each species assessed. We explore genetic associations with XCI by computing correlations between X-linked heterozygosity and XCI ratios, finding extremely low correlations, and by computing the strength of XCI ratios as a predictor for individual X-linked variants, revealing only a small collection of variants with moderate associations to large XCI ratios across the species. In summary, our work fails to find pervasive genetic correlates with XCI ratios and instead presents the inherent stochasticity of XCI as a general model for explaining population XCI ratio variability and inferring early developmental lineage decisions across mammalian species.

4. Preservation of co-expression defines the primary tissue fidelity of human neural organoids

4.1. Citation

Werner JM, Gillis J. 2023. Preservation of co-expression defines the primary tissue fidelity of human neural organoids. 2023.03.31.535112. https://www.biorxiv.org/content/10.1101/2023.03.31.535112v1 (Accessed May 1, 2023).

4.2. Author contributions and Acknowledgements

JMW and JG conceived the project. JMW and JG designed analyses. JMW performed analyses. JMW and JG wrote the manuscript.

JMW was supported by NSF award no. DGE-1938105. JG and JMW were supported by NIH grants R01MH113005 and R01LM012736. This material is based upon work supported by the National Science Foundation Graduate Research Fellowship Program under grant no. DGE-1938105. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation. We thank Tomasz Nowakowski and members of the Gillis lab for helpful comments on the manuscript.

4.3. Results summary

In this work, we compile a broad collection of scRNA-seq data from primary neural tissue and neural organoid datasets, spanning gestational weeks 5-25, sampling from 15 different brain regions and 12 different organoid differentiation protocols, for a total of 2.95 and 1.63 million cells for primary tissue and organoid datasets respectively. This broad sampling of primary tissue variation enables us to identify universal primary tissue signatures that constitute a generalizable benchmark for quantifying successes and failures of neural organoid systems. Specifically, quantifying the strength of preserved gene co-expression relationships across *in vivo* and *in vitro* datasets enables a functional, field-wide assessment of the current capabilities for modeling human brain development *in vitro*. We first compute meta-analytic primary tissue cell-type markers (MetaMarkers), deriving sets of markers that are expressed in a cell-type specific manner across all primary tissue datasets, timepoints and brain regions, constituting universal primary tissue cell-type signatures. We then construct co-expression networks from independent primary tissue data and the neural organoid datasets and quantify the strength of co-expression within the MetaMarker gene sets. Intra-gene set co-expression is high and comparable across all primary tissue data. Neural organoids exhibit extensive variance in performance, ranging from comparable to a complete lack of signal next to primary tissue data. We next measure the preservation of co-expression between primary tissue and organoid data within the MetaMarker gene sets, quantifying the degree the top-ten co-expressed partners of any given gene are preserved across systems. Using a meta-analytic primary tissue co-expression network as the reference, we demonstrate preservation of primary tissue co-expression is universally high across all primary tissue cell-types, with neural organoids again exhibiting high variability in performance. At a genome-wide level, we reveal that organoids consistently fail to preserve co-expression of ECM-related genes, suggesting ECM regulation is a strong and persistent failure of neural organoid models. In conclusion, we derive universal primary tissue cell-type signatures that are robust to temporal, regional, and technical variation, constituting a general benchmark for quantifying the fidelity of any neural organoid dataset. Specifically, our approach provides gene-, cell-type, and genome-wide quantification of the successes and failures of *in vitro* neural organoid differentiation. In line with the broader theme of developmental lineage for this thesis, this work presents a field-wide functional assessment of the current technical capabilities for modeling human neural development *in vitro*.

4.4 Introduction

Pluripotent stem cells create self-organized multi-cellular structures, termed organoids, when cultured in a 3D *in vitro* environment(Eiraku et al. 2008; Sato et al. 2009). The advantage of organoid models over 2D cell culture counterparts is their ability to generate structures that resemble endogenous tissues both in the differentiated cell-types produced and their 3D spatial organization(Lancaster et al. 2013; Corrò et al. 2020). The ability to model organogenesis in a controlled *in vitro* environment creates opportunities to study previously inaccessible developmental tissues from both humans and a range of model organisms(Pollen et al. 2019),(Kanton et al. 2019; Benito-Kwiecinski et al. 2021). As such, organoids are genetically accessible(Fleck et al. 2022) and environmentally perturbable(Sarieva and Mayer 2021) models enabling the study of molecular, cellular, and developmental mechanisms behind tissue construction. However, the applicability of studies in organoids to *in vivo* biology hinges on how well these *in vitro* models recapitulate primary tissue developmental processes, which remains an open question.

Quantifying the degree to which organoid systems replicate primary tissue biological processes is a critical step toward understanding the strengths and limitations of these *in vitro* models(Camp et al. 2015; Velasco et al. 2019; Bhaduri et al. 2020; Gordon et al. 2021; Feng et al. 2022). However, studies that perform such primary tissue/organoid comparisons are inherently confounded by batch(Leek et al. 2010) (*in vivo* vs *in vitro*), making it difficult to disentangle batch effects from underlying primary tissue and organoid biology. Meta-analytic approaches across many primary tissue and organoid datasets offer a route around these confounds, enabling the discovery of replicable primary tissue and organoid signatures independent of batch, which can then be interrogated for how well organoids recapitulate primary tissue biology(Tanaka et al. 2020; Cheroni et al. 2022; Kim et al. 2023). An important biological signature for this purpose is gene co-expression(Zhang and Horvath 2005). Genes that are functionally related tend to be expressed together, resulting in correlated gene expression dynamics that can define functionally relevant gene modules(Zhang and Horvath 2005). Gene co-expression relationships represent a shared genomic space that can be aggregated across experiments (e.g.,(Lee et al. 2020)) in either *in vivo* or *in vitro* systems, thus providing a useful framework for quantifying functional similarities and differences. Excitingly, coupling meta-analytic comparisons of primary tissue and organoid co-expression with single-cell RNA-sequencing data (scRNA-seq) stands to deliver cell-type specific quantifications of organoids' current capacity for producing functionally equivalent cell-types to primary tissues(Crow et al. 2016),(Mead et al. 2018).

Among organoid systems, human neural organoids are particularly well suited for meta-analytic evaluation due to well-described broad cell-type annotations and their known lineage relationships(Agboola et al. 2021), the wide variety of differentiation protocols in use(Mayhew and Singhania 2023), and the increasing amount of single-cell primary brain tissue and neural organoid data publicly available. In particular, the diversity of differentiation protocols for human neural organoids poses a unique challenge for organoid quality control that can be met by meta-analytic approaches. Neural organoids can either be undirected(Lancaster and Knoblich 2014) (multiple brain region identities) or directed (specific brain region identity) with an increasing number of protocols striving to produce a wider variety of region-specific organoids(Velasco et al. 2019; Muguruma et al. 2015; Sakaguchi et al. 2015; Qian et al. 2016; Xiang et al. 2017; Birey et al. 2017; Xiang et al. 2019; Miura et al. 2020; Eura et al. 2020; Andersen et al. 2020; Huang et al. 2021; Nayler et al. 2021; Sozzi et al. 2022). Meta-analytic primary tissue/organoid comparisons across differentiation protocols stand to derive generalizable quality control metrics applicable to any differentiation protocol, fulfilling a currently unmet need for unified quality control metrics across heterogeneous neural organoids.

Prior comparisons between primary brain tissues and neural organoids demonstrated that organoids have the capacity to produce diverse cell-types that capture both regional and temporal variation similar to primary tissue data as assayed through transcriptomic(Camp et al. 2015),(Velasco et al. 2019; Gordon et al. 2021; Tanaka et al. 2020; Cheroni et al. 2022), (Uzquiano et al. 2022), epigenomic(Luo et al. 2016; Amiri et al. 2018), electrophysiologic(Fair et al. 2020), and proteomic studies(Nascimento et al. 2019). At the morphological level, neural organoids can produce cellular organizations structurally similar to various *in vivo* brain regions, including cortical layers(Qian et al. 2020) and hippocampus(Sakaguchi et al. 2015), as well as modeling known inter-regional interactions like neuromuscular junctions(Andersen et al. 2020) and interneuron migration(Xiang et al. 2017). Additionally, several prior studies have compared primary tissue/organoid co-expression and concluded that neural organoids recapitulate primary brain tissue co-expression(Pollen et al. 2019; Gordon et al. 2021; Luo et al. 2016), but these assessments are highly targeted to study-specific properties, limiting potential generalization or potential assessment across the field. Typically, only a single organoid differentiation protocol is used in these assessments and it remains unclear whether organoids across different protocols will produce similar results. This lack of breadth also affects the use of primary tissue data used as a reference, with the primary tissue datasets utilized being treated as gold-standard datasets with little consideration for the extent one primary tissue reference may generalize to another. While prior meta-analytic comparisons of primary tissue/organoid co-expression have been performed(Cheroni et al. 2022), these were done at the bulk level (lack cell-type resolution) and included a small number of cortical organoid protocols, limiting the biological resolution and generalizability of these findings.

In this study, we perform a meta-analytic assessment of primary brain tissue (2.95 million cells, 50 datasets, Fig. 4.1A) and neural organoid (1.63 million cells, 130 datasets, 12 protocols, Fig. 4.1B) scRNA-seq datasets, constructing robust primary tissue cell-type specific markers and co-expression to query how well neural organoids recapitulate primary tissue cell-type specific biology. We sample primary brain tissue data over the first and second trimesters and across 15 different developmentally defined brain regions, extracting lists of cell-type markers that define broad primary tissue cell-type identity regardless of temporal, regional, or technical variation (Fig. 4.1A). We derive co-expression networks from individual primary tissue and organoid datasets as well as aggregate co-expression networks across datasets (Fig. 4.1C). From these networks, we assess the strength of co-expression within primary tissue cell-type marker sets as well as the preservation of co-expression patterns between primary tissue and organoid data (Fig. 4.1D-E). We also provide an R package to download our primary tissue reference co-expression network and assay new neural organoid data using simple, meaningful, and fast statistics (Fig. 4.1F). By constructing robust primary tissue cell-type representations through meta-analytic approaches, we demonstrate the preservation of primary tissue cell-type co-expression provides both specific and generalizable characterization of the primary tissue fidelity of human neural organoids.

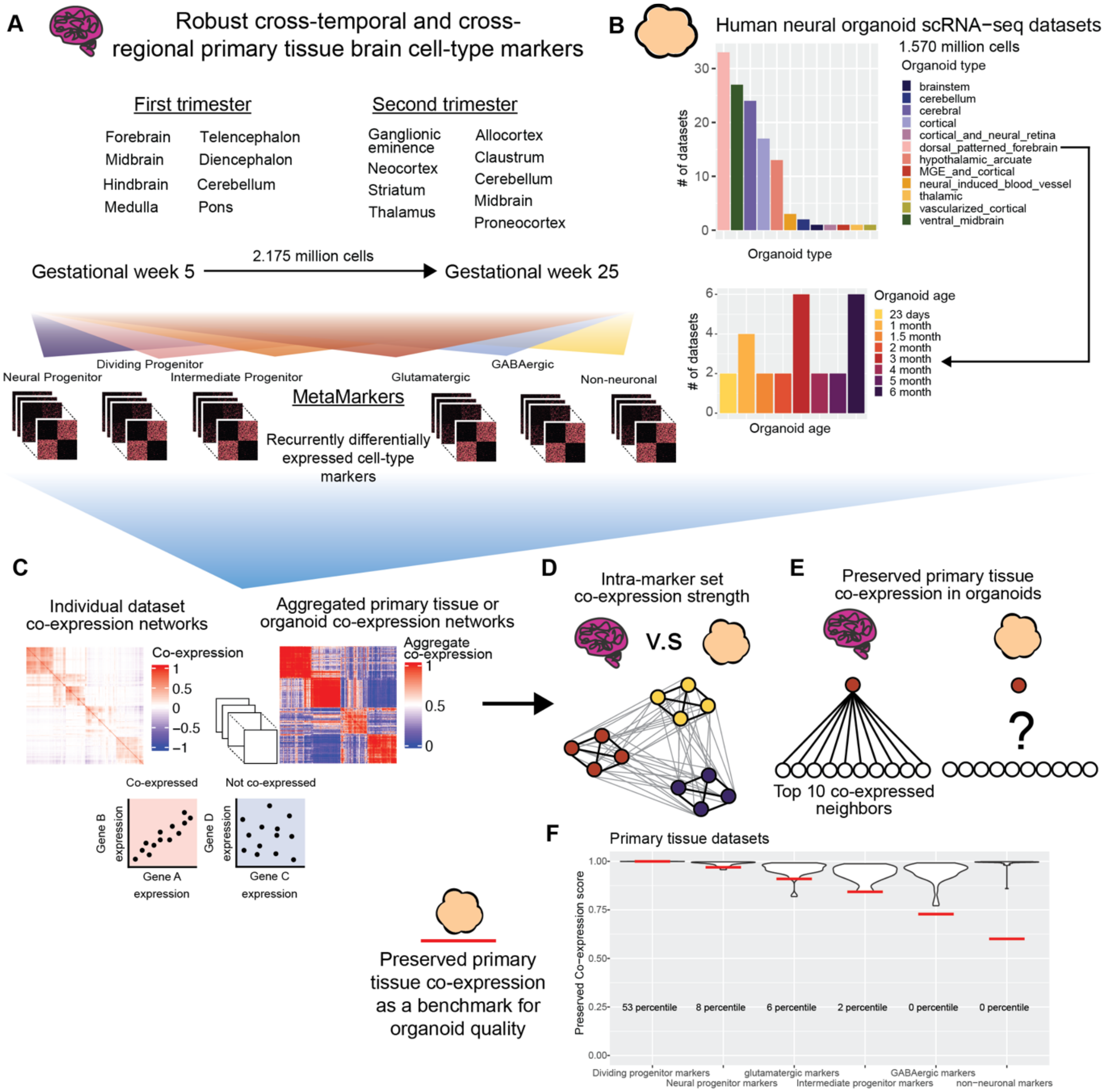

A
Robust cross-temporal and cross-regional primary tissue brain cell-type markers
First trimester
Forebrain
Midbrain
Hindbrain
Medulla
Telencephalon
Diencephalon
Cerebellum
Pons
Second trimester
Ganglionic eminence
Neocortex
Striatum
Thalamus
Allocortex
Claustrum
Cerebellum
Midbrain
Proneocortex
Gestational week 5
2.175 million cells
Gestational week 25
Neural Progenitor
Dividing Progenitor
Intermediate Progenitor
Glutamatergic
GABAergic
Non-neuronal
MetaMarkers
Recurrently differentially expressed cell-type markers
B
Human neural organoid scRNA−seq datasets
1.570 million cells
# of datasets
30
20
10
0
Organoid type
Organoid type
brainstem
cerebellum
cerebral
cortical
cortical_and_neural_retina
dorsal_patterned_forebrain
hypothalamic_arcuate
MGE_and_cortical
neural_induced_blood_vessel
thalamic
vascularized_cortical
ventral_midbrain
# of datasets
6
4
2
0
Organoid age
Organoid age
23 days
1 month
1.5 month
2 month
3 month
4 month
5 month
6 month
C
Individual dataset co-expression networks
Co-expression
1
0.5
0
−0.5
−1
Aggregated primary tissue or organoid co-expression networks
Aggregate co-expression
1
0.5
0
Co-expressed
Not co-expressed
Gene B expression
Gene A expression
Gene D expression
Gene C expression
D
Intra-marker set co-expression strength
V.S
E
Preserved primary tissue co-expression in organoids
?
Top 10 co-expressed neighbors
Preserved primary tissue co-expression as a benchmark for organoid quality
F
Primary tissue datasets
Preserved Co−expression score
1.00
0.75
0.50
0.25
0.00
53 percentile
8 percentile
6 percentile
2 percentile
0 percentile
0 percentile
Dividing progenitor markers
Neural progenitor markers
glutamatergic markers
Intermediate progenitor markers
GABAergic markers
non−neuronal markers

Figure 4.1: ***Using meta-analysis to quantify preserved primary tissue co-expression in organoids***
A Collection of annotated primary tissue brain scRNA-seq datasets, ranging from gestational week (GW) 5 to 25 and sampling from 15 developmentally defined brain regions. The primary tissue datasets are annotated at broad cell-type levels (Neural Progenitor, Dividing Progenitor, Intermediate Progenitor, Glutamatergic, GABAergic, and Non-neuronal) and these annotations are used to compute MetaMarkers, cell-type markers identified through recurrent differential expression.
B Collection of human neural organoid scRNA-seq datasets, sampling from 12 different differentiation protocols. Included is an annotated temporal forebrain organoid dataset.
C Example of a sparse co-expression network derived from a scRNA-seq data and an example of an aggregate co-expression network averaged over many scRNA-seq datasets. The aggregate network enhances the sparse signal from the individual network.
D Schematic showing a quantification of intra-marker set co-expression
E Schematic showing a quantification for the strength of preserved co-expression between two co-expression networks, measuring the replication of the top 10 co-expressed partners of an individual gene across the networks.
F Example plot from the preservedCoexp R library, placing cell-type specific preserved co-expression scores of an example forebrain organoid dataset in reference to scores derived from primary tissue datasets. Red lines denote the percentile of the organoid cell-type scores within the primary tissue distributions.

4.5. Results

4.5.1. Meta-analytic framework for primary tissue/organoid comparisons

We reason that, if they exist, primary tissue cell-type specific signals robust to temporal, regional, and technical variation will constitute *in vivo* standards applicable to any organoid dataset regardless of time in culture or differentiation protocol. We first show it is possible to learn sets of marker genes that define broad primary tissue cell-types (Fig. 4.1A, Supp. Table 4.1) across timepoints (gestational weeks GW5-GW25) and brain regions (15 developmentally defined brain regions) through a meta-analytic differential expression framework (Fig. 4.1A, Fig. 4.2A-B). We then compare co-expression within these marker sets between primary tissue and organoid data to quantify the degree organoids preserve primary tissue cell-type specific co-expression. An important aspect of our analysis is our cross-validation of primary tissue differential expression and co-expression. We employ a leave-one-out cross-validation approach when learning robust differentially expressed marker genes from our annotated primary tissue datasets (2,174,934 cells, 37 datasets) and we interrogate co-expression of our primary tissue marker genes within a large cohort of unannotated primary tissue datasets (776,343 cells, 14 datasets). This approach ensures we are extracting primary tissue markers and co-expression relationships independent of temporal, regional, and technical variation, a powerful approach for deriving broad primary tissue signatures appropriate for comparison to a wide range of organoid datasets.

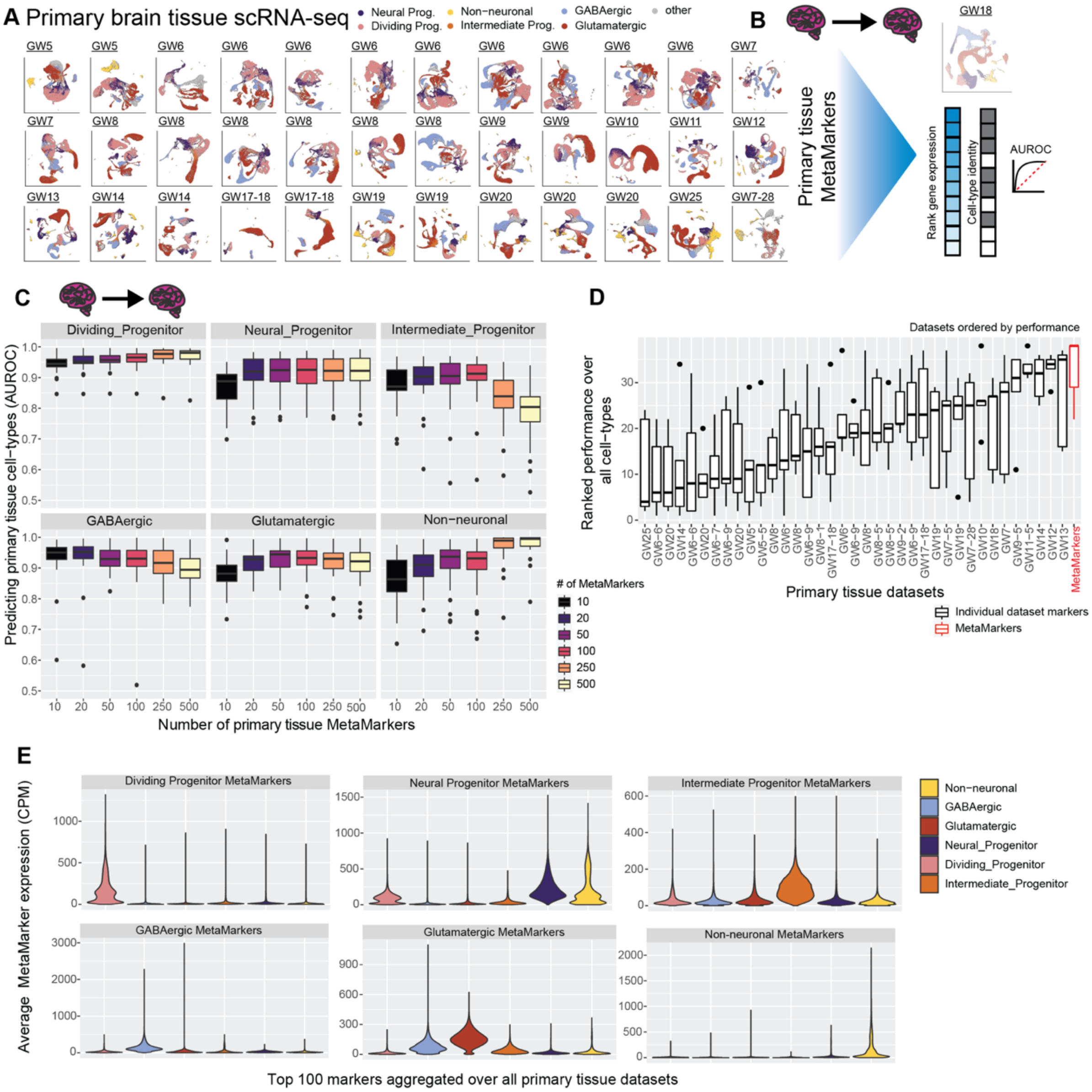


Figure 4.2: ***Meta-analytic primary tissue cell-type markers***
A Annotated UMAPs of the annotated primary tissue brain scRNA-seq datasets.
B Example of our leave-one-out cross-validation approach for learning primary tissue MetaMarkers and testing the markers' capacity for predicting annotations in the left-out dataset, quantified with the AUROC statistic.
C Meta-analytic primary tissue markers have high performance in predicting primary tissue cell-type annotations. Boxplot distributions of the AUROC statistic for predicting cell-type annotations across all leave-one-out combinations of our annotated primary tissue datasets, with an increasing number of MetaMarkers used for predicting cell-type annotations on the x-axis.
D MetaMarkers have the highest performance in predicting primary tissue cell-type annotations. Boxplots of marker gene-set performances. Gene-sets are the top 100 cell-type markers from individual primary tissue datasets compared to the MetaMarker performance. Performances for each cell-type in individual primary tissue datasets are presented in Supp. Fig. 4.1A. Datasets are ordered by their median performance.
E Averaged distributions of gene expression for the top 100 MetaMarkers demonstrating clear cell-type specificity. This is performed with a leave-one-out cross-validation, with individual dataset distributions reported in Supp. Fig. 4.1B.

4.5.2. Cross-temporal and -regional primary tissue cell-type markers

To learn markers that define broad primary tissue cell-types, we apply the MetaMarkers(Fischer and Gillis 2021) framework to our cross-temporal and -regional annotated primary tissue datasets (Fig. 4.2A-B). MetaMarkers uses robust differential expression statistic thresholds (log2 fold-change >= 4 and FDR-adjusted p-value <= 0.05) for determining whether a gene is differentially expressed (DE) within individual datasets, then ranks all genes via the strength of their recurrent DE across datasets (see methods). We test the generalizability of our primary tissue MetaMarker gene sets in predicting primary cell-types by employing a leave-one-out primary tissue cross-validation (Fig. 4.2A-B). We construct an aggregate expression predictor in the left-out dataset using MetaMarkers learned from the remaining datasets (see methods), quantifying how well the MetaMarker gene sets predict the left-out cell-type annotations with the area-under-the-receiver-operating-characteristic curve statistic (AUROC, Fig. 4.2B-C). The AUROC is the probability of correctly prioritizing a true positive (e.g., cell of the right type) above a negative, (e.g., cell of the wrong type), given some predictor of the positive class, in this case, aggregate cell-type marker expression.

Starting with just the top 10 primary tissue MetaMarkers per cell-type, we achieve a mean AUROC across all primary tissue datasets of 0.944 ± 0.0280 SD, 0.865 ± 0.0653 SD, 0.873 ± 0.0676 SD, 0.937 ± 0.0669 SD, 0.879 ± 0.0535 SD, and 0.863 ± 0.0768 SD, for dividing progenitors, neural progenitors, intermediate progenitors, GABAergic neurons, glutamatergic neurons, and non-neuronal cell-types respectively (Fig. 4.2C). These extremely high performances demonstrate that even a small number of meta-analytically derived primary tissue cell-type markers have high utility in predicting primary tissue cell-type annotations regardless of temporal and regional variability. For all following analysis, we take the top 100 MetaMarkers per cell-type as robust representations of our 6 broad primary tissue cell-type annotations (average AUROC >= 0.90 except for intermediate progenitors: 0.897 ± 0.0777 SD), with the 100 MetaMarkers achieving modest increases in performance over the top 10 MetaMarkers for all cell-types except GABAergic cells (Fig. 4.2C, mean AUROC for 100 GABAergic MetaMarkers: 0.922 ± 0.0777 SD). When comparing MetaMarkers to markers derived from individual primary tissue datasets, we find the MetaMarkers are consistently top performers in predicting primary tissue annotations (Fig. 4.2D), with MetaMarkers producing the top results for intermediate progenitors, glutamatergic neurons, and GABAergic neurons (Supp. Fig. 4.1), as well as comparable performance to top individual datasets for dividing progenitors, neural progenitors, and non-neuronal cell-types (Supp. Fig. 4.1).

We explore the primary tissue MetaMarker sets further by computing the average expression of the top 100 MetaMarkers for our 6 annotated cell-types across all cells within our 37 annotated primary tissue datasets (Fig. 4.2E), continuing our leave-one-out approach. Each annotated primary tissue cell-type expresses the corresponding matched MetaMarker set over all other MetaMarker sets, with the exception of some off-target expression for the neural progenitor MetaMarkers in dividing progenitors and non-neurons (aggregated over all datasets Fig. 4.2E, individual datasets Supp. Fig. 4.1B). This demonstrates our MetaMarker gene sets act as robust cell-type markers in aggregate across all first and second trimester timepoints (Fig. 4.2E, Supp. Fig. 4.1B). Additionally, we investigate the expression of the top 100 MetaMarker gene sets across annotated primary brain regions, demonstrating each primary tissue cell-type maximally expresses the corresponding primary tissue MetaMarker set across all annotated brain regions (Supp. Fig. 4.2A-B). Overall, we are able to meta-analytically extract cell-type markers that define broad primary tissue cell-types independent of temporal and regional variation.

4.5.3. Broad primary tissue cell-type markers capture organoid temporal variation

After extracting meta-analytic cell-type markers that capture broad primary tissue temporal and regional variation, we can test how well these markers also capture organoid temporal and regional (protocol) variation. We start with a large-scale temporal organoid atlas(Uzquiano et al. 2022) derived from a forebrain differentiation protocol containing timepoints ranging from 23 days to 6 months in culture. When comparing primary tissue and organoid data along a temporal axis, one might expect younger primary tissue expression data to be a better reference for younger organoid cell-types (better able to predict cell-types) and vice-versa for older primary and organoid data (Supp. Fig. 4.3A). We test this relationship using the same AUROC quantification as in Figure 4.1C, but now using the top 100 primary tissue cell-type markers per primary tissue dataset to predict organoid cell-type annotations across all organoid timepoints (Supp. Fig. 4.3B, see methods).

We observe highly consistent performance across all primary tissue datasets (GW5 – GW25) when predicting organoid cell-types regardless of the organoid timepoint (Supp. Fig. 4.3B). The average difference in AUROC scores when predicting organoid cell-types using either our youngest (GW5) or oldest (GW25) primary data is 0.000382 ± 0.0357 SD, 0.132 ± 0.188 SD, 0.141 ± 0.0980 SD, 0.0379 ± 0.130 SD and 0.0845 ± 0.209 SD for dividing progenitors, neural progenitors, glutamatergic neurons GABAergic neurons, and non-neuronal cells respectively (No annotated intermediate progenitors in the GW25 primary tissue dataset). This demonstrates strikingly consistent performance across primary tissue timepoints, highlighting that broad primary tissue cell-type signatures are applicable as reference for organoid cell-types regardless of the primary tissue or organoid timepoint. The one exception is for neural progenitors, where there seemingly is a temporal shift in performance with younger primary tissue datasets predicting younger organoid annotations over older organoid annotations and vice-versa for older primary tissue/organoid data (Supp. Fig. 4.3B). However, a subset of the young GW6-8 primary tissue datasets report sharp increases in performance predicting older organoid timepoints in opposition to other GW6-8 primary tissue datasets, suggesting variance in performance is driven by intersections between the quality of individual organoid and primary tissue datasets rather than overarching temporal variability. Importantly, our lists of top 100 primary tissue MetaMarkers perform comparably to marker sets from individual primary tissue datasets, with less variance in performance across the organoid timepoints for the differentiated cell-types (mean AUROC variance across organoid timepoints for individual primary tissue datasets vs. primary MetaMarker variance; glutamatergic: 0.0142, 0.00477, GABAergic: 0.00921, 0.00199, non-neuronal: 0.00901, 0.00670, Supp. Fig. 4.3B). This demonstrates our meta-analytic primary tissue cell-type markers robustly capture organoid temporal variation.

4.5.4. Broad primary tissue cell-type markers capture organoid protocol variation

We assess whether our primary tissue MetaMarker gene sets capture organoid variation outside the annotated forebrain temporal organoid atlas by performing principal-component analysis (PCA) across all organoid datasets, representing data from 12 different differentiation protocols. Our lists of 100 primary tissue MetaMarkers are consistently heavily weighted in the first PC across organoid datasets (Supp. Fig. 4.3C-D). While a large portion of PC1-weighted genes are dividing progenitor MetaMarkers (representing cell-cycle signal), markers for non-dividing fetal cell-types also comprise those genes consistently heavily weighted in PC1 across organoid datasets (Supp. Fig. 4.3C-D).

4.5.5. Aggregate organoid co-expression preserves primary tissue cell-type co-expression

Our primary tissue MetaMarkers that capture both primary tissue and organoid temporal/regional variation enable assessments of cell-type specific co-expression between arbitrary primary tissue and organoid datasets. One normally would need matched cell-type annotations across datasets to compare cell-type specific biology, but here we couple our meta-analytically derived cell-type markers with gene co-expression quantifications, which do not rely on cell-type annotations, to extract cell-type specific co-expression from any given scRNA-seq dataset. Practically, if organoids are producing cell-types functionally identical to primary tissue cell-types, we would expect near identical co-expression relationships within our primary tissue MetaMarker gene sets across primary tissue and organoid datasets.

Deriving co-expression relationships from single-cell data is challenging due to inherent sparsity of the expression data (Fig. 4.3A). We overcome this sparsity with straightforward standardization and aggregation approaches (Fig. 4.3A, see methods), which prioritize replicable signal across datasets. We first explore marker set co-expression within our unannotated primary tissue datasets, which were not included in deriving our primary tissue MetaMarker sets. The aggregate unannotated primary tissue co-expression network nearly perfectly constructs cell-type specific co-expression modules when hierarchically clustering the co-expression of our top 100 primary tissue MetaMarker gene sets (Fig. 4.3B).

Turning to the aggregated organoid co-expression network, while the intra- and inter-MetaMarker gene set co-expression appears dysregulated compared to the unannotated primary tissue co-expression network, the overall clustering of MetaMarker genes by co-expression still largely captures cell-type specific clustering (Fig. 4.3B). We quantify this through the Adjusted Rands Index (ARI) metric, comparing the MetaMarker clustering through co-expression in any given network to the perfect clustering of MetaMarker gene sets by cell-type. We perform this quantification for both the aggregated co-expression networks (diamond, triangle, and square special characters, Fig. 4.3C) and for all individual primary tissue and organoid co-expression networks (boxplots, Fig. 4.3C). While individual organoid networks perform worse than individual primary tissue networks on average, the aggregated organoid network is largely comparable to individual primary tissue networks (Fig. 4.3C, median annotated and unannotated primary tissue network ARI: 0.403, 0.437, aggregated organoid network ARI: 0.381). In aggregate, organoid co-expression largely captures broad primary tissue cell-type specific co-expression.

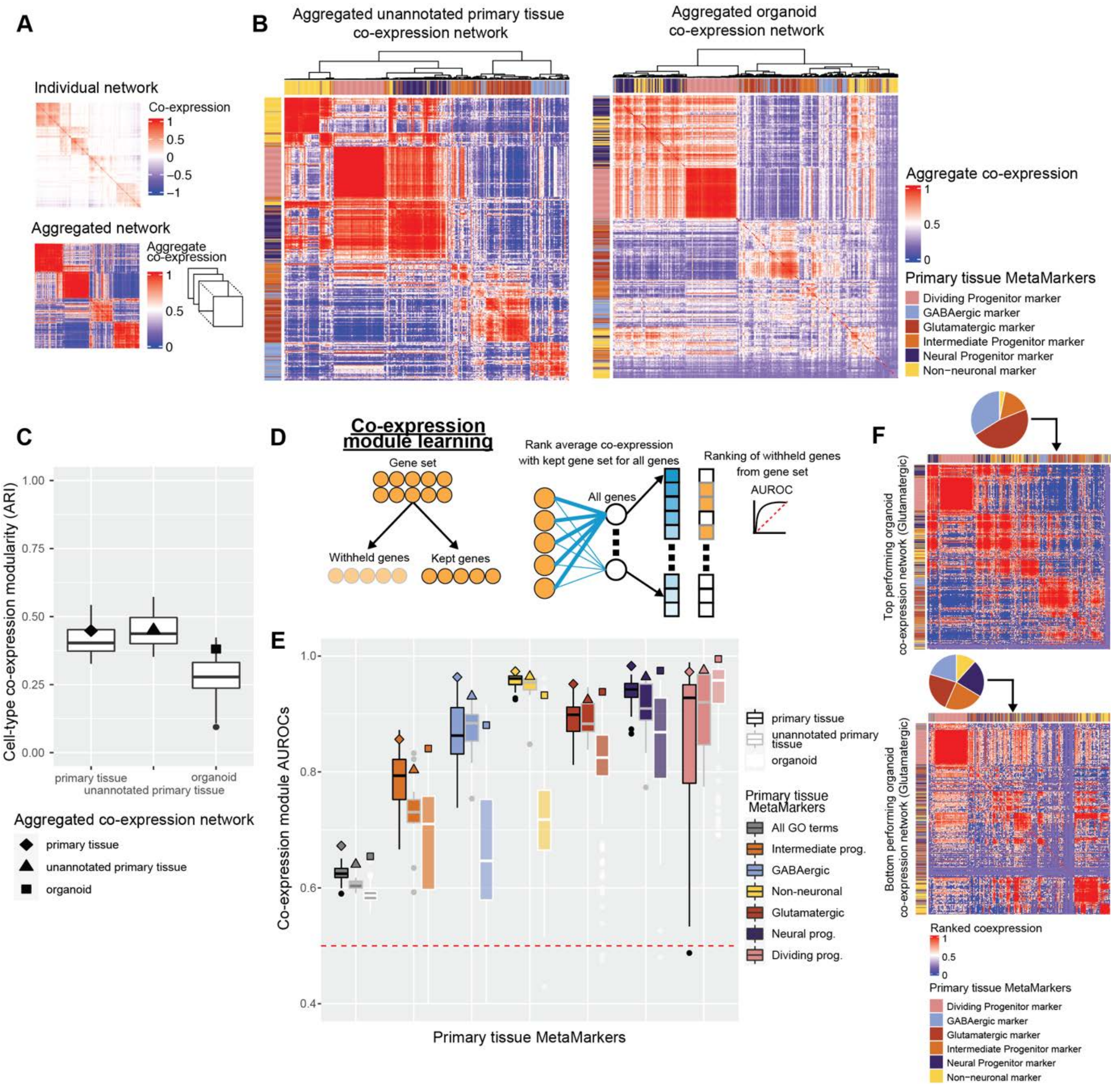

A
Individual network
Co-expression
1
0.5
0
−0.5
−1
Aggregated network
Aggregate co-expression
1
0.5
0
B
Aggregated unannotated primary tissue co-expression network
Aggregated organoid co-expression network
Aggregate co-expression
1
0.5
0
Primary tissue MetaMarkers
Dividing Progenitor marker
GABAergic marker
Glutamatergic marker
Intermediate Progenitor marker
Neural Progenitor marker
Non−neuronal marker
C
Cell-type co-expression modularity (ARI)
1.00
0.75
0.50
0.25
0.00
primary tissue
unannotated primary tissue
organoid
Aggregated co-expression network
◆ primary tissue
▲ unannotated primary tissue
■ organoid
D
Co-expression module learning
Gene set
Withheld genes
Kept genes
Rank average co-expression with kept gene set for all genes
All genes
Ranking of withheld genes from gene set
AUROC
E
Co-expression module AUROCs
1.0
0.8
0.6
0.4
Primary tissue MetaMarkers
primary tissue
unannotated primary tissue
organoid
Primary tissue MetaMarkers
All GO terms
Intermediate prog.
GABAergic
Non-neuronal
Glutamatergic
Neural prog.
Dividing prog.
F
Top performing organoid co-expression network (Glutamatergic)
Bottom performing organoid co-expression network (Glutamatergic)
Ranked coexpression
1
0.5
0
Primary tissue MetaMarkers
Dividing Progenitor marker
GABAergic marker
Glutamatergic marker
Intermediate Progenitor marker
Neural Progenitor marker
Non−neuronal marker

Figure 4.3: ***Neural organoids vary in recapitulating primary tissue cell-type marker set co-expression***
A Example of a sparse co-expression network derived from a scRNA-seq data and an example of an aggregate co-expression network averaged over many scRNA-seq datasets. The aggregate network enhances the sparse signal from the individual network.
B Marker gene-sets show clear cell-type clusters via their co-expression relationships in primary tissue and organoid networks. The aggregated co-expression networks for the unannotated primary tissue datasets and organoid datasets, showing the hierarchically clustered co-expression of the primary tissue MetaMarkers for the 6 cell-types.
C Organoid cell-type clustering via co-expression is notable lower compared to all primary tissue datasets. Distributions of the Adjusted Rands Index (ARI) for individual annotated primary tissue, unannotated primary tissue, and organoid datasets. The ARI scores for the aggregate networks are denoted with the special characters.
D Schematic for the co-expression module learning framework, measuring the co-expression strength within an arbitrary gene-set compared to the rest of the genome, quantified with the AUROC statistic.
E Distributions of co-expression module AUROCs for individual annotated primary tissue, unannotated primary tissue, and organoid datasets for the co-expression strength of the MetaMarker gene-sets for the 6 cell-types. The grey ‘All GO terms’ distributions report the average co-expression module AUROC across all GO terms for each individual dataset. Co-expression module AUROCs for the aggregate co-expression networks are denoted with the special characters.
F Top and bottom organoid co-expression networks based on Glutamatergic performance. Heatmaps depict the hierarchically clustered co-expression of the primary tissue MetaMarker gene-sets for the 6 cell-types. Cell-type specific clusters are apparent in the top network, but are more mixed in the bottom network. Pie-charts depict the percentage of MetaMarker gene-sets that make up an example cluster in each network determined via the hierarchical clustering.

4.5.6. Organoid datasets vary in primary tissue cell-type marker set co-expression

Having broadly assessed co-expression across our MetaMarker gene sets, we then asked how well do organoids recapitulate primary tissue co-expression within each cell-type specific MetaMarker gene set. We score intra-gene set co-expression strength through a simple machine learning framework(Ballouz et al. 2017),(Skinnider et al. 2019), which quantifies whether genes in a given set are more strongly co-expressed with each other compared to the rest of the genome (Fig. 4.3D).

Co-expression module scores across the annotated and unannotated primary tissue datasets are largely comparable with the exception of a sharp decrease in intermediate progenitor performance for the unannotated primary tissue datasets (Fig. 4.3E). Six out of the fourteen unannotated datasets are sampled from either the ganglionic eminences or the hypothalamus, potentially explaining this decrease in performance and suggesting our intermediate progenitor MetaMarkers are enriched for signal from cortical areas. In contrast, performance is much more variable across the individual organoid datasets for all cell-types except the dividing progenitors, ranging from no signal (AUROC <= 0.50) to comparable results with primary tissue networks (Fig. 4.3E). We visualize the top and bottom performing organoid co-expression networks for glutamatergic co-expression to highlight the extreme variability across organoid datasets in recapitulating primary tissue co-expression (Fig. 4.3F). In the top performing organoid network, we find cell-type specific co-expression modules with a clear glutamatergic module (Fig. 4.3F). While co-expression for dividing progenitor markers constructs a clear module in the bottom performing organoid network, non-dividing primary tissue cell-type co-expression is clearly dysregulated with clusters composed of all primary tissue cell-type markers (Fig. 4.3F). By quantifying the intra-gene set co-expression of our primary tissue MetaMarkers, we are able to place organoid datasets on a spectrum ranging from complete failure to primary tissue-level recapitulation of primary tissue co-expression.

Importantly, organoid datasets vary strongly by protocol type in recapitulating primary tissue cell-type specific co-expression (Supp. Fig. 4.4). The undirected differentiation protocols (cerebral and cortical, Supp. Fig. 4.4) produce highly variable results across the primary tissue cell-types, in line with previous reports of high variability for undirected organoids. Intriguingly, the vascularized cortical organoid protocol produces consistently high performance across all primary tissue cell-types (Supp. Fig. 4.4), suggesting vascularized models increase organoids' capacity to produce comparable primary tissue cell-types *in vitro*. We also find the vascularized cortical, dorsal patterned forebrain, and undirected cortical protocols produce some of the highest co-expression module scores for our GABAergic primary tissue MetaMarkers (Supp. Fig. 4.4). These results agree with previous observations of *in vitro* production of inhibitory cell-types within these cortical models previously expected to produce exclusively excitatory lineages.

4.5.7. Organoid datasets vary in preserving gene-level primary tissue co-expression

We take our primary tissue/organoid co-expression comparisons a step further and ask how well individual organoid datasets preserve gene-level primary tissue co-expression relationships. For any given individual gene, we can quantify whether that gene's top co-expressed partners are preserved in one co-expression network compared to another (Fig. 4.4A). We use the aggregated co-expression network from the annotated primary tissue datasets as our reference co-expression network and test how well individual co-expression networks, either primary tissue or organoid, perform in preserving primary tissue gene-level co-expression patterns (Fig. 4.4A, top 10 co-expressed neighbors). We start by quantifying the preserved co-expression of genes within our primary tissue MetaMarker gene sets, using the average preserved co-expression AUROC as a measure of preserved co-expression for any given gene set (Fig. 4.4A). Across our 6 annotated primary tissue cell-types, primary tissue co-expression networks deliver consistently high performance for preserved co-expression scores of our primary tissue MetaMarker gene sets (Fig. 4.4B, mean preserved co-expression score across cell-types and primary tissue datasets: annotated 0.971 ± 0.0227 SD, unannotated 0.963 ± 0.00957 SD). This indicates the top 10 co-expressed partners are highly preserved for the vast majority of genes within each MetaMarker gene set across all primary tissue datasets.

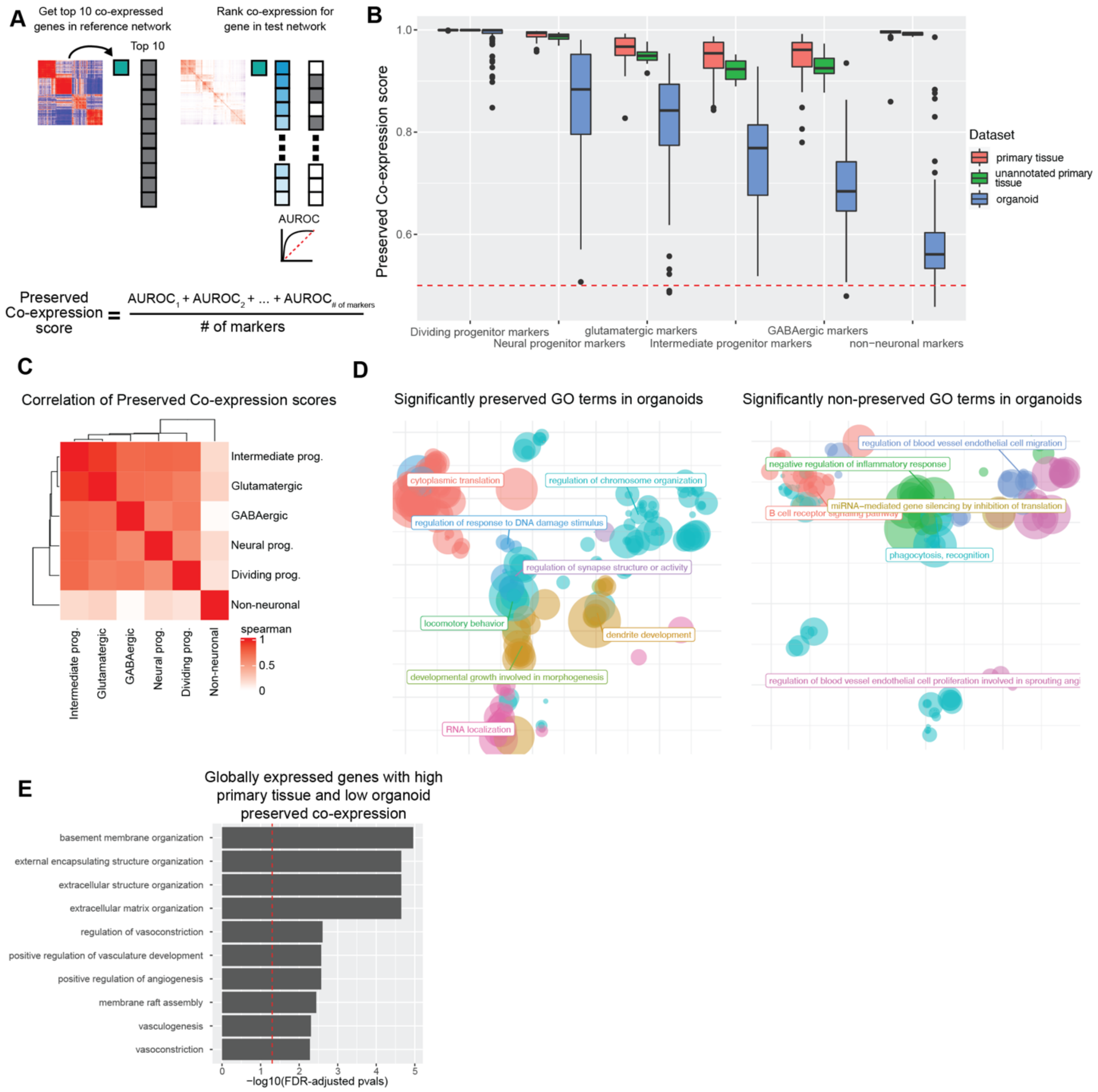

A
Get top 10 co-expressed genes in reference network
Rank co-expression for gene in test network
Top 10
AUROC
Preserved Co-expression score = (AUROC_1 + AUROC_2 + ... + AUROC_# of markers) / # of markers
B
1.0
0.8
0.6
Preserved Co-expression score
Dividing progenitor markers
Neural progenitor markers
glutamatergic markers
Intermediate progenitor markers
GABAergic markers
non−neuronal markers
Dataset
primary tissue
unannotated primary tissue
organoid
C
Correlation of Preserved Co-expression scores
Intermediate prog.
Glutamatergic
GABAergic
Neural prog.
Dividing prog.
Non-neuronal
spearman
1
0.5
0
D
Significantly preserved GO terms in organoids
cytoplasmic translation
regulation of chromosome organization
regulation of response to DNA damage stimulus
regulation of synapse structure or activity
locomotory behavior
dendrite development
developmental growth involved in morphogenesis
RNA localization
Significantly non-preserved GO terms in organoids
regulation of blood vessel endothelial cell migration
negative regulation of inflammatory response
miRNA−mediated gene silencing by inhibition of translation
B cell receptor signaling pathway
phagocytosis, recognition
regulation of blood vessel endothelial cell proliferation involved in sprouting angi
E
Globally expressed genes with high primary tissue and low organoid preserved co-expression
basement membrane organization
external encapsulating structure organization
extracellular structure organization
extracellular matrix organization
regulation of vasoconstriction
positive regulation of vasculature development
positive regulation of angiogenesis
membrane raft assembly
vasculogenesis
vasoconstriction
0
1
2
3
4
5
−log10(FDR-adjusted pvals)

Figure 4.4: ***Neural organoids vary in their preservation of primary tissue gene-level co-expression***
A Schematic showing the quantification for gene-level preserved co-expression. The preserved co-expression score for any given gene-set is the average preserved co-expression AUROC across all genes within that gene set.
B Organoids strongly vary in preserved primary tissue cell-type specific co-expression in comparison to fetal data. Boxplot distributions show the preserved co-expression scores for the primary tissue MetaMarker gene-sets of the 6 cell-type annotations across all individual networks.
C The majority of cell-types are significantly correlated in preserved co-expression within organoid networks. Spearman correlation matrix for the preserved co-expression scores for all 6 cell-type annotations across all individual organoid datasets.
D Scatter plots summarizing the semantic distances of GO terms that are significantly preserved or non-preserved between the aggregate annotated primary tissue and organoid co-expression networks.
E Organoids globally fail to preserve primary tissue co-expression of ECM and vascular related genes. Bar plot detailing the top 10 GO terms from a GO enrichment test of the 76 genes with high and low preserved co-expression AUROCs within primary tissue networks and organoid networks respectively. The preserved co-expression for each individual gene from primary tissue networks and organoid networks is reported in Supp. Fig. 4.6E.

In contrast, individual organoid datasets vary substantially in preserved co-expression scores across our primary tissue MetaMarker gene sets (Fig. 4.4B). As before with our quantification of intra-gene set co-expression, quantifying preserved gene-level co-expression places organoid datasets on a spectrum of near zero to indistinguishable preserved co-expression to primary tissue data. Organoid datasets vary substantially by protocol in preserving primary tissue cell-type specific co-expression, echoing similar trends as observed from our co-expression module analysis (Supp. Fig. 4.5). Since the majority of our organoid protocols are designed for producing excitatory lineages, it is encouraging we report a higher average preservation of glutamatergic primary tissue co-expression over non-neuronal or GABAergic primary tissue co-expression across our organoid datasets (Fig. 4.4B). Unsurprisingly, preservation of dividing progenitor co-expression is universally high with a preserved co-expression score of approximately 1 in nearly every primary tissue and organoid dataset, representing consistent co-expression of cell-cycle marker genes across systems (Fig. 4.4B, Supp. Fig. 4.5). A subset of organoid datasets are clear outliers to this trend (Fig. 4.4B, Supp. Fig. 4.5), suggesting that cell-cycle co-expression is not preserved, indicating basic cellular functions may be dysregulated in these datasets. One intriguing observation came from a study that compared organoids grown in a vertical shaker versus an orbital shaker(Suong et al. 2021). We show that organoids grown in an orbital shaker produce higher preserved primary tissue co-expression scores for intermediate progenitors and glutamatergic cell-types whereas organoids grown in a vertical shaker produce higher scores for GABAergic cell-types (3 replicates each, glutamatergic, intermediate progenitor, GABAergic; Orbital: 0.896 ± 0.00102 SD, 0.795 ± 0.00148 SD, 0.665 ± 0.0308 SD. Vertical: 0.644 ± 0.0125 SD, 0.686 ± 0.0161 SD, 0.763 ± 0.00731 SD). This suggests the mechanical conditions of organoid growth can distinctly impact lineage and cell-type production in organoids.

With measures of preserved primary tissue co-expression for multiple cell-types within organoids, we can additionally assess variation in preserved co-expression across cell-types within individual organoid datasets. We compute correlations of preserved co-expression scores between the 6 MetaMarker sets across all organoid datasets and find significantly positive correlations (FDR-adjusted p-value < .001) across all comparisons with the exception of the non-neuronal cell-type (Fig. 4.4C, non-neuronal FDR-adjusted p-values range from < 0.001 to 0.745). This indicates preserved primary tissue co-expression is a global feature of organoid datasets. For example, if an organoid is producing neural progenitors that preserve primary tissue co-expression, that organoid is likely producing other cell-types that preserve primary tissue co-expression. Similarly, we asked if preserved co-expression varies across normal or perturbed organoids. A subset of our organoid datasets come from studies that performed various perturbations (22q11.2 deletion, SMARCB1 knockdown, exposure to Alzheimer's serum, SETBP1 point mutations, amyotrophic lateral sclerosis patient-derived organoids). We compare the MetaMarker preserved co-expression scores between normal and perturbed organoids and find only a single significant difference across all cell-type MetaMarker sets (intermediate progenitor normal vs. mutant preserved co-expression score FDR-adjusted p-value: 0.0287, Supp. Fig. 4.6). This demonstrates our broad primary tissue cell-type co-expression signatures are also applicable for comparison with organoids in perturbation experiments.

After revealing cell-type specific variation for preserving primary tissue co-expression within organoids, our co-expression networks additionally allow genome-wide assessments of preserved co-expression. We extend our analysis via GO terms to quantify preserved primary tissue co-expression within organoids across the whole genome. GO terms with significantly preserved primary tissue co-expression (see methods) in organoids are mostly related to basic cellular functions like response to DNA damage and protein translation, as well as GO terms related to neurodevelopment (Fig. 4.4D). GO terms that significantly lack preservation of primary tissue co-expression are almost exclusively related to angiogenesis or immune function (Fig. 4.4D), concordant with the fact that organoids lack vasculature and an immune system.

While GO terms are useful for partitioning the genome into functional units for comparison, our co-expression networks also enable assessments of preserved co-expression for individual genes. As a particular use-case, we search for genes with exceptionally high preserved primary tissue co-expression across primary tissue datasets that also have poor preserved primary tissue co-expression across organoid datasets. We only consider genes that have some measurable expression in every organoid and primary tissue dataset and compute the average preserved co-expression AUROC for each gene across the organoid and primary tissue datasets (Supp. Fig. 4.6). The top 10 enriched GO terms for genes (76 in total) with high primary tissue (average AUROC >= 0.99) and low organoid (average AUROC < 0.70) preserved co-expression are related to extra-cellular matrix (ECM) and vascular characterizations (Fig. 4.4E). The poor conservation of genes related to vasculature can be explained by the absence of vascularization in the vast majority of our organoid datasets. The subset of these 76 genes in the ECM GO terms are CAV1, CAV2, COL4A1, CTSK, ENG, LAMB1, LAMC1, NID1, NID2, DDR2, and VWA1. Notably, these genes produce collagen and laminins, components of Matrigel, the artificial ECM typically included in organoid cultures. These results highlight preserved primary tissue co-expression of ECM-related genes as a particularly consistent deficit across neural organoids, suggesting that investigations into the signaling between artificial ECM and cells in organoid cultures may be a route forward for general improvements of organoid fidelity.

In summary, we interrogate co-expression in organoids at multiple levels, revealing organoids vary in preserving primary tissue co-expression at gene-, cell-type, and whole genome resolutions through the use of a robust aggregate primary tissue co-expression network.

4.5.8 Temporal variation in organoid preservation of primary tissue co-expression

We score preserved co-expression in organoids using the aggregate primary tissue co-expression network (GW5-25), which by design aims to capture signal robust to temporal variation. To investigate temporal trends in organoid co-expression, we employ a similar approach as when predicting organoid cell-type annotations (Supp. Fig. 4.3), this time quantifying the preservation of primary tissue co-expression for the top 100 cell-type markers per individual primary tissue dataset across all organoid timepoints (Fig. 4.5A-B). We uncover a broad temporal shift in the preservation of primary tissue co-expression within organoids across all cell-types, with younger organoids (23 days – 1.5 months) as the top performers for mostly first trimester primary tissue co-expression transitioning to older organoids (2 – 6 months) as top performers for mostly second trimester primary tissue co-expression (Fig. 4.5B). This temporal shift is broadly consistent across the cell-types, beginning around GW9-10 (Fig. 4.5B). Our approach in predicting organoid annotations in Figure 4.2 is based on aggregate marker expression and did not produce temporally variable results, whereas our approach here comparing preserved co-expression of the same marker genes does produce temporally variable results. This indicates that the co-expression relationships of genes rather than their expression levels better capture temporal variation in developing systems.

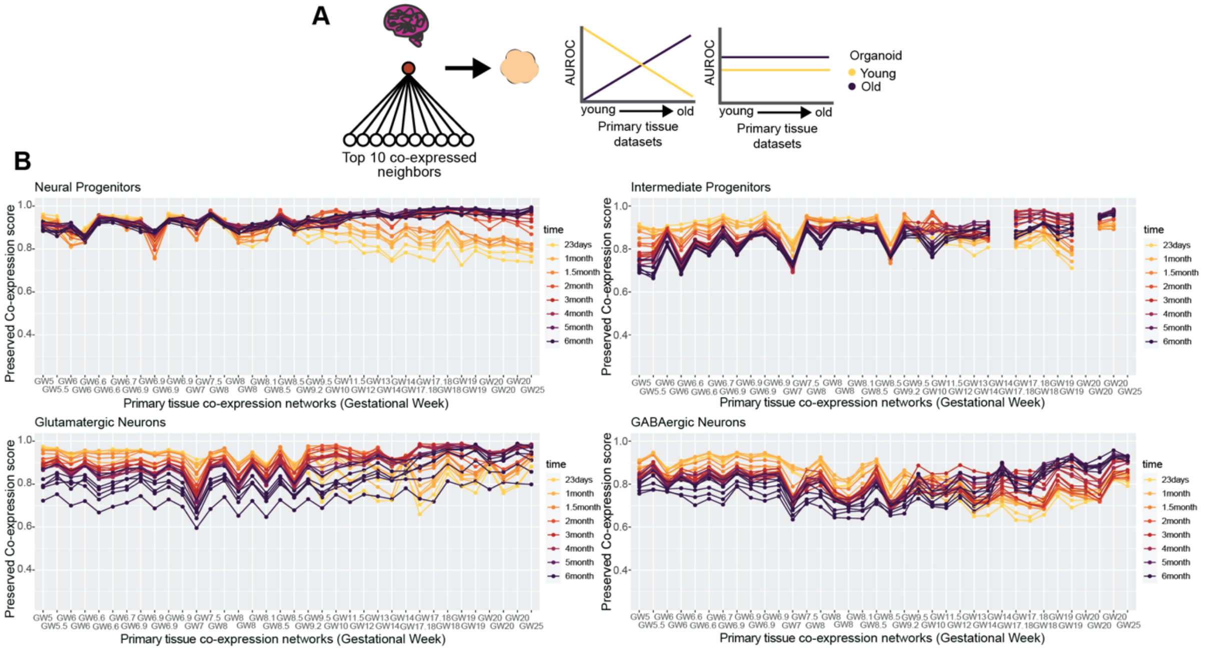


Figure 4.5: ***Neural organoids capture temporal dynamics in primary tissue co-expression***
A Schematic showing two potential outcomes when comparing the preserved co-expression between primary tissue and organoid data on a temporal axis. There may be a temporal relationship, with younger organoids recapitulating younger primary tissue co-expression over older primary tissue co-expression and vice versa for older organoids, or there may be no temporal relationship.
B Organoid co-expression models temporal trends in primary tissue co-expression. Line plots showing the preserved co-expression scores computed from individual organoid co-expression networks for cell-type markers of individual primary tissue datasets. Primary tissue datasets on the x-axis are ordered from youngest to oldest.

4.5.9. Organoids preserve developing brain co-expression over adult brain co-expression

We demonstrate temporal variation in developing brain co-expression relationships is captured by organoids, but only from the single forebrain organoid protocol used in the temporal organoid atlas. In order to extend analysis across all our organoid datasets and assess broad temporal variation in co-expression, we next investigate the preserved co-expression within organoids of both developing and adult brain co-expression relationships.

We construct an aggregate adult co-expression network from a medial temporal gyrus scRNA-seq dataset(Jorstad et al. 2022) (157,508 cells) sampling 7 adult individuals. We compare the preserved co-expression scores of organoids for either developing or adult glutamatergic, GABAergic, and non-neuronal cell-types. Organoids unanimously preserve developing brain co-expression over adult co-expression (Supp. Fig. 4.6) for glutamatergic and GABAergic cell-types with equally poor performance for the non-neuronal cell-type, suggesting organoids generally fail to produce non-neuronal cell-types. We extend this analysis genome-wide and place organoids in context between developing and adult data by computing the average preservation of co-expression AUROC across all genes for organoid, developing, and adult co-expression using the annotated primary developing brain tissue network as the reference. The adult co-expression network produces a global preserved developing brain co-expression score of 0.591, indicating very poor performance across the genome in preserving developing co-expression relationships (Supp. Fig. 4.6). Organoids vary substantially in their global preservation of developing brain co-expression with some organoid datasets performing comparably to the adult data. This result is largely influenced by the number of cells present within individual organoid datasets (Supp. Fig. 4.6, corr 0.647, p-value < .001), suggesting a cell-sampling limitation for uncovering developing brain co-expression within organoids. However, organoid datasets report more variable global preserved co-expression scores compared to down-sampled developing brain data (Supp. Fig. 4.6), indicating a remaining biological gap between primary developing brain tissue and organoid data not explained through technical means.

An intriguing study generated data from human cortical organoids either transplanted or not into developing rat brains to test the limits of maturation organoids can achieve *in vitro*(Revah et al. 2022). We compare the preservation of developing and adult co-expression between these age-matched non-transplanted and transplanted human cortical organoids. We report that while the non-transplanted organoids preserve developing co-expression over adult for glutamatergic and GABAergic markers (Supp. Fig. 4.6, non-transplanted glutamatergic and GABAergic mean developing brain AUROCs: 0.797 ± 0.0281 SD, 0.697 ± 0.0212 SD. Non-transplanted glutamatergic and GABAergic mean adult AUROCs: 0.672 ± 0.0234 SD, 0.586 ± 0.0309 SD), the transplanted organoids have increased preservation of adult co-expression for glutamatergic and non-neuronal cell-types (Supp. Fig. 4.6, transplanted glutamatergic and non-neuronal mean developing brain AUROCs: 0.759 ± 0.00908 SD, 0.501 ± 0.0127 SD. Transplanted glutamatergic and non-neuronal mean adult AUROCs: 0.849 ± 0.0332 SD, 0.738 ± 0.00779 SD). This indicates the transplanted human organoids are adopting adult human glutamatergic and non-neuronal co-expression, concordant with the original authors' conclusions of increased maturation in transplanted organoids.

4.5.10. Variability in organoid co-expression is driven by marker gene expression

We investigate the impact of various technical features in our analysis on our co-expression results by assessing their correlation with our co-expression module scores and preserved co-expression AUROCs, focusing on technical features like sequencing depth, number of cells, etc. An important technical consideration for our analysis is ensuring all datasets have an identical gene namespace for meaningful comparisons of expression data. We fit all datasets to the GO gene universe, dropping gene annotations not in GO or zero-padding missing GO annotations in individual datasets. Excessive zero-padding of genes within our MetaMarker gene sets may artificially lower co-expression module scores or preserved co-expression scores, though we find this relationship to be relatively weak with little impact on score variance (Supp. Fig. 4.7, R^2 for co-expression module scores and zero-padding: 0.00123, 0.0165, 0.103, 0.0284, 0.0337, 0.052, R^2 for preserved co-expression and zero-padding: 0.0825, 0.321, 0.149, 0.0302, 0.0409, 0.000769 for neural prog., dividing prog., intermediate prog., glutamatergic, GABAergic, and non-neuronal cell-types respectively). Sequencing depth is also similarly found to have little impact on our co-expression module scores or preserved co-expression scores (Supp. Fig. 4.7). Rather, the features strongly related to performance are the number of cells in a dataset and the strength of marker set expression, all significantly strongly correlated with co-expression module scores and preserved co-expression scores (Supp. Fig. 4.7, range of significant (p-value < .001) correlations between marker set expression or cell number and co-expression module scores or preserved co-expression scores: 0.351 – 0.808, with the exception of dividing progenitors having a significant negative correlation of -0.453 between marker set expression and co-expression module scores).

4.5.11. Preservation of primary tissue co-expression as a generalizable quality control metric

As a general summary, our approach for quantifying preserved primary tissue co-expression across numerous organoid protocols revealed the axes on which organoids lie for recapitulating primary tissue co-expression relationships at gene, cell-type, and whole-genome resolutions. These assessments provide powerful quality control information, identifying which genes and/or cell-types organoids can or cannot currently model on par with primary tissue data. We make our methods accessible through an R package to aid in future organoid studies and protocol development, providing means for rapidly constructing co-expression networks from scRNA-seq data (Fig. 4.6A) as well as querying preserved co-expression of users' data with our aggregate primary tissue brain co-expression network (Fig.

4.6A). Additionally, we make the results of our meta-analysis across primary tissue and organoid datasets available for users to place their data in reference to a field-wide collection (Fig. 4.6B).

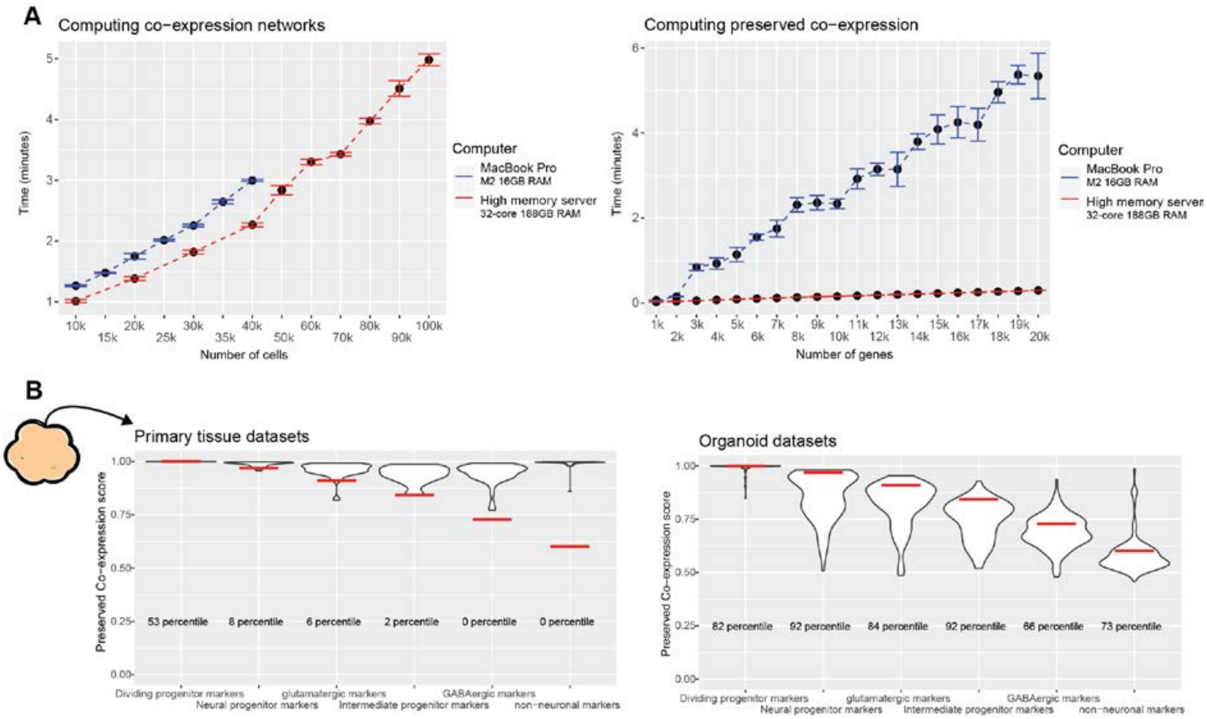


Figure 4.6: ***The preservedCoexp R package enables fast computation of preserved co-expression***
A The preservedCoexp R package can compute co-expression networks and genome-wide preservation of co-expression in a few minutes even for low-memory computers. Line plots showing the computational time to either compute co-expression networks or preserved co-expression as the number of cells or genes increases. Points are the mean value from 10 replicates, with error bars depicting ± 1 standard deviation.
B Example plot from the preservedCoexp R package, placing cell-type specific preserved co-expression scores of an example forebrain organoid dataset in reference to scores derived from primary tissue datasets or organoid datasets. Red lines denote the percentile of the forebrain organoid cell-type scores within either the primary tissue distributions or organoid distributions.

4.6. Methods

4.6.1 Dataset download and scRNA-seq pre-processing

Links for all downloaded data (GEO accession numbers, data repositories, etc.) are provided in Supp. Table 1. All scRNA-seq data was processed using the Seurat v4.2.0 R package. Data made available in 10XGenomics format (barcodes.tsv.gz, features.tsv.gz, matrix.mtx.gz) were converted into Seurat objects using the Read10X() and CreateSeuratObject() Seurat functions. Data made available as expression matrices were converted into sparse matrices and then converted into Seurat objects using the CreateSeuratObject() function. Ensembl gene IDs were converted into gene names using the biomaRt v2.52.0(Durinck et al. 2009) package.

Where metadata was made available, we separated data by batch (Age, Donor, Cell line, etc.) for our final total of 130 organoid and 51 primary tissue datasets (Supp. Table 1). We processed and analyzed each batch independently without integration. We used consistent thresholds for filtering cells across all datasets, keeping cells that had less than 50% of reads mapping to mitochondrial genes and had between 200 and 6000 detected genes. Several datasets provided annotations for potential doublets; we excluded all cells labeled as doublets when annotations were made available. All data made available with raw expression counts were CPM normalized with NormalizeData(normalization.method = 'RC', scale.factor = 1e6), otherwise normalizations were kept as author supplied.

For primary tissue and organoid data made available with cell-type annotations, we provide our mapping between author provided annotations and our broad cell-type annotations in Supp. Table 2.

4.6.2. Primary tissue MetaMarker generation and cross-validation

MetaMarkers were computed using the MetaMarkers v0.0.1(Fischer and Gillis 2021) R package, which requires shared cell-type and gene annotations across datasets to derive a ranked list of MetaMarkers. Gene markers for individual datasets were first computed using the compute_markers() function on the CPM normalized expression data for our annotated primary tissue datasets (Supp. Table. 1). A ranked list of MetaMarkers was then computed using the make_meta_markers() function using all 37 individual annotated primary tissue dataset marker lists. Genes are first ranked through their recurrent differential expression (the number of datasets that gene was called as DE using a threshold of log2 FC >= 4 and FDR p-value <= .05) and then through the averaged differential expression statistics of each gene across individual datasets. When we take the top 100 markers per individual dataset as in Fig. 4.2D, Fig. 4.5, Supp. Fig. 4.1A, and Supp. Fig. 4.3B, we rank markers for each dataset by their AUROC statistic as computed with the compute_markers() MetaMarkers function.

For the cross-validation of our primary tissue MetaMarkers, we excluded a single annotated primary tissue dataset, computed MetaMarkers from the remaining 36 annotated primary tissue datasets, and then used those MetaMarkers to predict the cell-type annotations of the left-out dataset. We construct an aggregate expression predictor to quantify the predictive strength a list of genes has, in this case our MetaMarker lists, in predicting cell-type annotations. Taking any arbitrary number of genes (10, 20, 50, 100, 250, or 500 MetaMarkers), we sum the expression counts for those genes within each cell and then rank all cells by this aggregate expression vector. We compute an AUROC using this ranking and the cell-type annotations for a particular cell-type through the Mann-Whitney U test. Formally:

$$AUROC = \frac{U}{n_0 * n_1}$$

where U is the Mann-Whitney U test statistic, n_0 is the number of positives (cells with a given cell-type annotation) and n_1 is the number of negatives (cells without that cell-type annotation).

$$U = R_0 - \frac{n_0(n_0 + 1)}{2}$$

where R_0 is the sum of the positive ranks.

As an example, if there are 10 genes that are perfect glutamatergic markers (only glutamatergic cells express these genes), then ranking cells by the summed expression of these genes will place all glutamatergic cells (positives) in front of all other cells (negatives), producing an AUROC of 1. The violin plots in Supp. Fig. 4.1B and in Figure 4.2E visualize our aggregate expression approach, where datapoints per cell-type are the aggregated expression counts for the given top 100 MetaMarkers across all cells per dataset (Supp. Fig. 4.1B) or aggregated across all datasets (Fig. 4.2E).

For Supp. Fig. 4.1A, we took the top 100 cell-type markers per individual primary tissue dataset (x-axis) and used those genes to predict cell-type annotations as described above for all other annotated primary tissue datasets, reported as the AUROC boxplot distributions. The MetaMarker distribution was computed using a leave-one-out approach as described above. We ranked the individual primary tissue datasets by their median AUROC performance per cell-type to derive the distributions of ranks presented in Figure 4.2D, excluding the dividing progenitor data as performance was highly consistent across all primary tissue datasets.

4.6.3. Cross-regional primary tissue MetaMarker expression

We investigated the aggregate expression of our top 100 MetaMarkers per cell-type across annotated brain regions separately for the annotated first-trimester and second-trimester primary tissue atlases due to differing regional annotations. MetaMarkers were computed with a leave-one-out approach as described above using all 37 of the annotated primary tissue datasets. For the heatmaps in Supp. Fig. 4.2, rows represent the annotated cells present within the given dataset, columns represent the aggregated expression for the top 100 given cell-type MetaMarkers and each annotated region present. We average the aggregated expression for each cell-type per region and then normalize each region (column) by the maximum average expression value across the cell-types. A value of 1 indicates that cell-type is the one maximally expressing the given MetaMarker set for that brain region. The heatmaps are ordered by cell-type and region and are not clustered.

4.6.4. Organoid PCA

PCA analysis was performed using the Seurat function RunPCA() with the top 2000 variable features, determined using the Seurat function FindVariableFeatures(selection.method = 'vst', nfeatures = 2000). For each organoid dataset, we took the eigenvector for the first principal component, computed the absolute value, and then divided by the maximum value to compute a normalized vector between 0 and 1. We visualized the normalized eigenvectors for each organoid dataset in Supp. Fig. 4.3C, keeping primary tissue MetaMarker genes that were detected in the top 2000 variable genes of at least 10 organoid datasets. Genes missing from any given dataset's top 2000 variable genes were given a value of 0. The heatmap was produced using the ComplexHeatmap v2.12.1(Gu et al. 2016) package and was hierarchically clustered using the ward.D2 method for both rows and columns.

4.6.5. Generating co-expression networks from scRNA-seq data

To generate a shared gene annotation space across all datasets, we fit each dataset to the GO gene universe before computing co-expression matrices. Using human GO annotations (sourced 2022-03-10 using the org.Hs.eg.db v3.15.0(Carlson 2019) and

AnnotationDbi v1.58.0(Pagès et al. 2022) R packages), we excluded gene expression from a dataset if the gene annotation was not present in GO and we zero-padded missing GO genes for each dataset.

We compute a gene-by-gene co-expression matrix per dataset using the spearman correlation coefficient computed across all cells in a given dataset. We then rank the correlation coefficients in the gene-by-gene matrix and divide by the maximum rank to obtain a rank-standardized co-expression matrix. All results reported using individual dataset co-expression networks (Fig. 4.3C-F, Fig. 4.4B, Figs. 4.5-4.6, Supp. Figs. 4.4-4.7) were obtained using the rank-standardized co-expression networks.

We compute the aggregated co-expression networks by taking the average of the rank standardized co-expression networks for each gene-gene index.

4.6.6. Hierarchical clustering of primary tissue MetaMarkers by co-expression

We visualize the co-expression of primary tissue MetaMarker genes using the ComplexHeatmap package and the ward.D2 algorithm for hierarchical clustering. We use the fossil v0.4.0 package(Vavrek 2020) to compute the adjusted Rands Index with the adj.rand.index() function. To compute the adjusted Rands Index, we calculate a consensus clustering of MetaMarkers per co-expression network across 100 k-means clusterings (using the arguments row_km = 6, column_km = 6, row_km_repeats = 100, column_km_repeats = 100 within the Heatmap function) to compare to the perfect grouping of MetaMarkers by cell-type.

4.6.7. Co-expression module learning analysis

EGAD v1.24.0(Ballouz et al. 2017) is a machine learning framework that quantifies the strength of co-expression within an arbitrary gene-set compared to the rest of the genome with an AUROC quantification (Fig. 4.3D). We compute co-expression module AUROCs for all GO gene-sets (between 10 and 1000 genes per GO term) and our top 100 primary tissue MetaMarker gene-sets for each individual primary tissue and organoid co-expression network as well as the aggregated annotated, unannotated and organoid networks. For the annotated primary tissue co-expression networks, we employ a leave-one-out approach, learning MetaMarkers from 36 of the annotated datasets and computing co-expression module AUROCs for these MetaMarkers in the left-out dataset's co-expression network. We compute co-expression module AUROCs using the EGAD run_GBA() function with default parameters. In Figure 4.3E, the 'All GO terms' distributions report the average co-expression module AUROC across all GO terms for each individual network.

4.6.8. Preservation of co-expression

To compute our preservation of co-expression AUROC, we take the top 10 co-expressed partners for gene A in a reference co-expression network as our positive gene annotations. In a test co-expression network, we rank all genes through their co-expression with gene A and compute an AUROC using this ranking and the positive annotations derived from the reference network. If gene A in the test network has the exact same top 10 co-expressed partners as in the reference network, that would result in an AUROC of 1. To summarize a given gene-set's preserved co-expression, we take the average preserved co-expression AUROC across all genes in that gene set as the preservation of co-expression score for that gene set. We use the aggregated annotated primary tissue co-expression matrix as our reference network.

The preserved co-expression scores for the annotated primary tissue data in Figure 4.4B were computed with a leave-one-out approach. MetaMarkers and an aggregated co-expression matrix were computed from 36 of the annotated primary tissue datasets and then preserved co-expression scores were computed using the co-expression network of the left-out annotated primary tissue dataset.

4.6.9. Preservation of GO term co-expression

We compute p-values for the preservation of co-expression of GO terms using a mean sample error approach. Using the aggregated annotated primary tissue co-expression network as the reference and the aggregated organoid network as the test network, we first compute the preserved co-expression AUROCs for all individual genes, taking the mean and standard deviation value as the population mean and population standard deviation. For any given GO term, we first compute the preserved co-expression score for the term (the average of the preserved co-expression AUROCs for the genes in the term) and then compute the sample error for that score with:

$$SE = \frac{SD_{pop}}{\sqrt{n_{GO}}}$$

where $SD_{pop}$is the population standard deviation and n_{GO} is the number of genes in the GO term. We then compute a z-score through:

$$Z_{GO} = \frac{mu_{GO} - mu_{pop}}{SE}$$

where mu_{go} is the preserved co-expression score for the GO term and mu_{pop} is the population mean preserved co-expression AUROC. We compute left-sided p-values using the standard normal distribution:

$$p_L = P(X \leq Z_{GO})$$

Where X is a normal distribution with mean = 0 and standard deviation = 1. We use the R function pnorm(Z_{GO}) to compute this p-value.

We then compute the right-sided p-value as:

$$p_R = 1 - p_L$$

We adjust p-values using the R function p.adjust(method = ‘BH’). We filter for GO terms that have between 20 and 250 genes per term and use a threshold of FDR-corrected p-value <= 0.0001 to call significance. Significant left-sided p-values are interpreted as GO terms with significantly smaller preserved co-expression scores (significantly not preserved) than expected through sampling error and right-sided p-values are interpreted as GO terms with significantly larger preserved co-expression scores (significantly preserved) than expected through sampling error. We use the R package rrvgo to visualize the significant GO terms in Fig. 4.4D.

4.6.10. Computing correlation significance

We employ a permutation test to compute p-values for any given correlation coefficient. We permute data-pairs and compute a correlation coefficient, repeating for 10,000 random permutations to generate a distribution of correlation coefficients under the null hypothesis of independence. We calculate a two-sided p-value for the original correlation coefficient as the number of permuted correlation coefficients whose absolute value is greater than or equal to the absolute value of the original correlation coefficient, divided by 10,000. We adjust p-values using the R function p.adjust(method = 'BH') and use a FDR-corrected p-value threshold of <= .05 to call significance.

4.6.11. Comparing co-expression of normal vs. perturbed organoids

For both the co-expression module AUROCs and the preserved co-expression scores of normal and perturbed organoids, we test for significant differences per cell-type using the Mann Whitney U test, adjusting p-values with the R function p.adjust(method = 'BH') and using a FDR-corrected p-value threshold of <= .05 to call significance.

4.6.12. Organoid temporal analysis

The organoid temporal analysis for both predicting organoid annotations with primary tissue markers (Supp. Fig. 4.3B) and scoring the preserved co-expression of organoid co-expression using primary tissue networks as reference (Fig. 4.5) were performed for all pair-wise combinations of the 37 annotated primary tissue datasets and the 26 temporally annotated forebrain organoid datasets. We excluded the GW7-28 annotated primary tissue dataset from the temporal preserved co-expression analysis (Fig. 4.5) due to the wide temporal range sampled. For predicting organoid annotations with primary tissue markers, we used the top 100 markers per primary tissue dataset to construct aggregate expression predictors in the organoid datasets as described above. The MetaMarkers performance was calculated using MetaMarkers derived from all 37 annotated primary tissue datasets. For scoring preserved co-expression, individual primary tissue networks were used as the reference with individual organoid networks as the test networks. We computed the preserved co-expression scores of the top 100 primary tissue cell-type markers per individual primary dataset for each individual organoid network.

4.6.13. GO enrichment analysis

We compute enrichment for GO terms using Fisher's Exact Test as implemented through the hypergeometric test. We compute raw p-values for GO terms with between 10-1000 genes and compute FDR-adjusted p-values using p.adjust(method = 'BH'). We only consider GO sets with between 20 and 500 when choosing the top 10 GO sets in Figure 4.4E, ranked by FDR-adjusted p-value.

4.6.14. R and R packages

All analysis was carried out in R v4.2.2. Colors with selected using the MetBrewer v0.2.0 R library. Plots were generated using ggplot2 v3.3.6(Wickham 2016). Spearman correlation matrices for co-expression networks were computed using a python v3.6.8 script, implemented in R with the reticulate v1.26 R package, as well as using functions from the matrixStats v0.62.0 R library. All code used in generating results and visualizations will be made public at the time of publication. The preservedCoexp R library is made available at https://github.com/JonathanMWerner/preservedCoexp.

4.7. Supplemental Figures

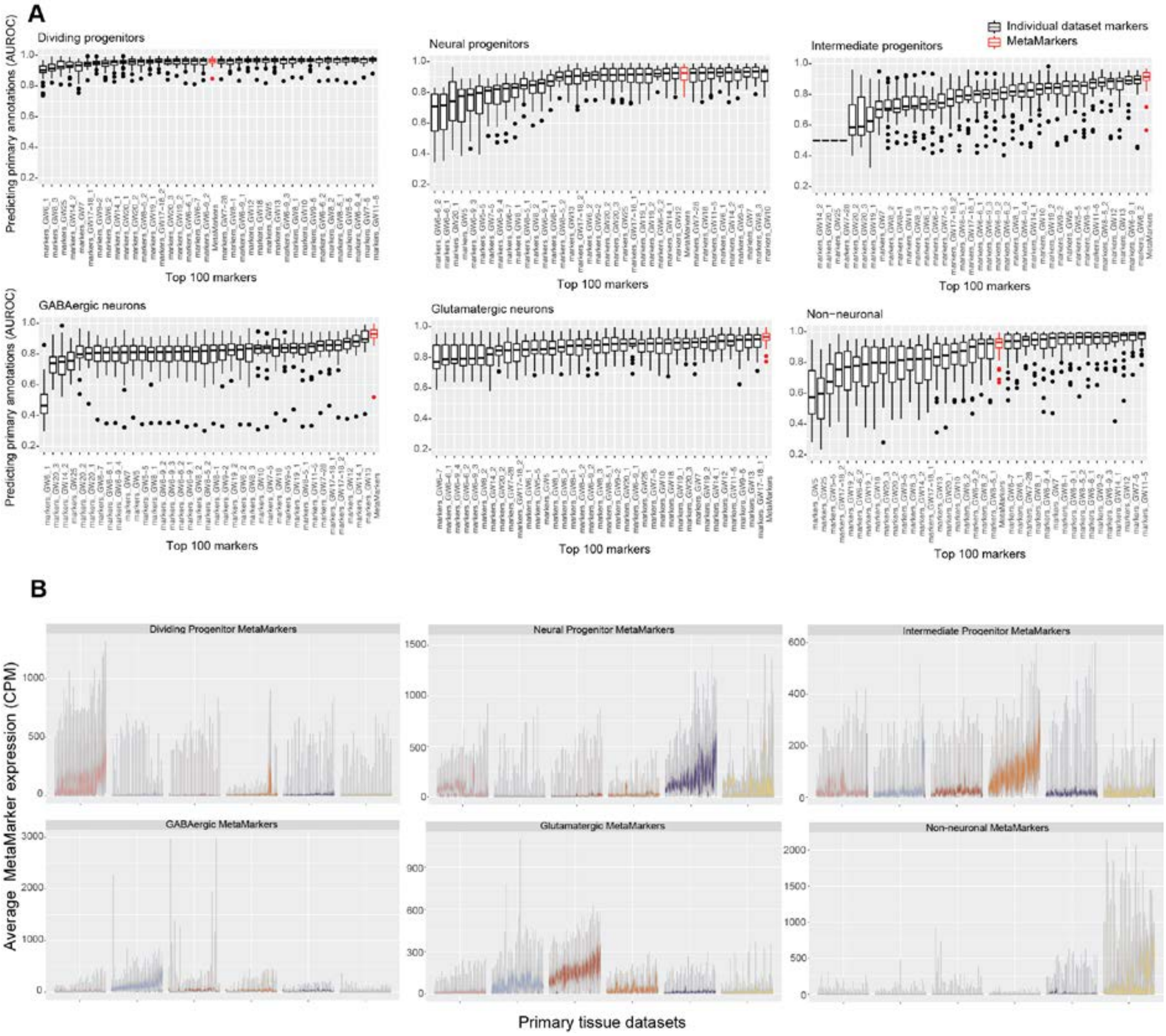


Figure S4.1 MetaMarkers as temporally robust primary tissue cell-type markers
A MetaMarkers are consistent top performers in predicting primary tissue cell-type annotations. Boxplots of AUROCs for predicting cell-type annotations across all primary tissue datasets using the top 100 marker genes per individual primary tissue dataset compared to MetaMarkers (red). Datasets are ordered by their median performance, providing the rank distributions in Figure 2D.
B MetaMarkers exhibit cell-type specificity across all primary tissue datasets. Averaged distributions of gene expression for the top 100 MetaMarkers across all annotated primary tissue datasets with leave-one-out cross-validation. Figure 2E is the aggregate over these individual dataset distributions.

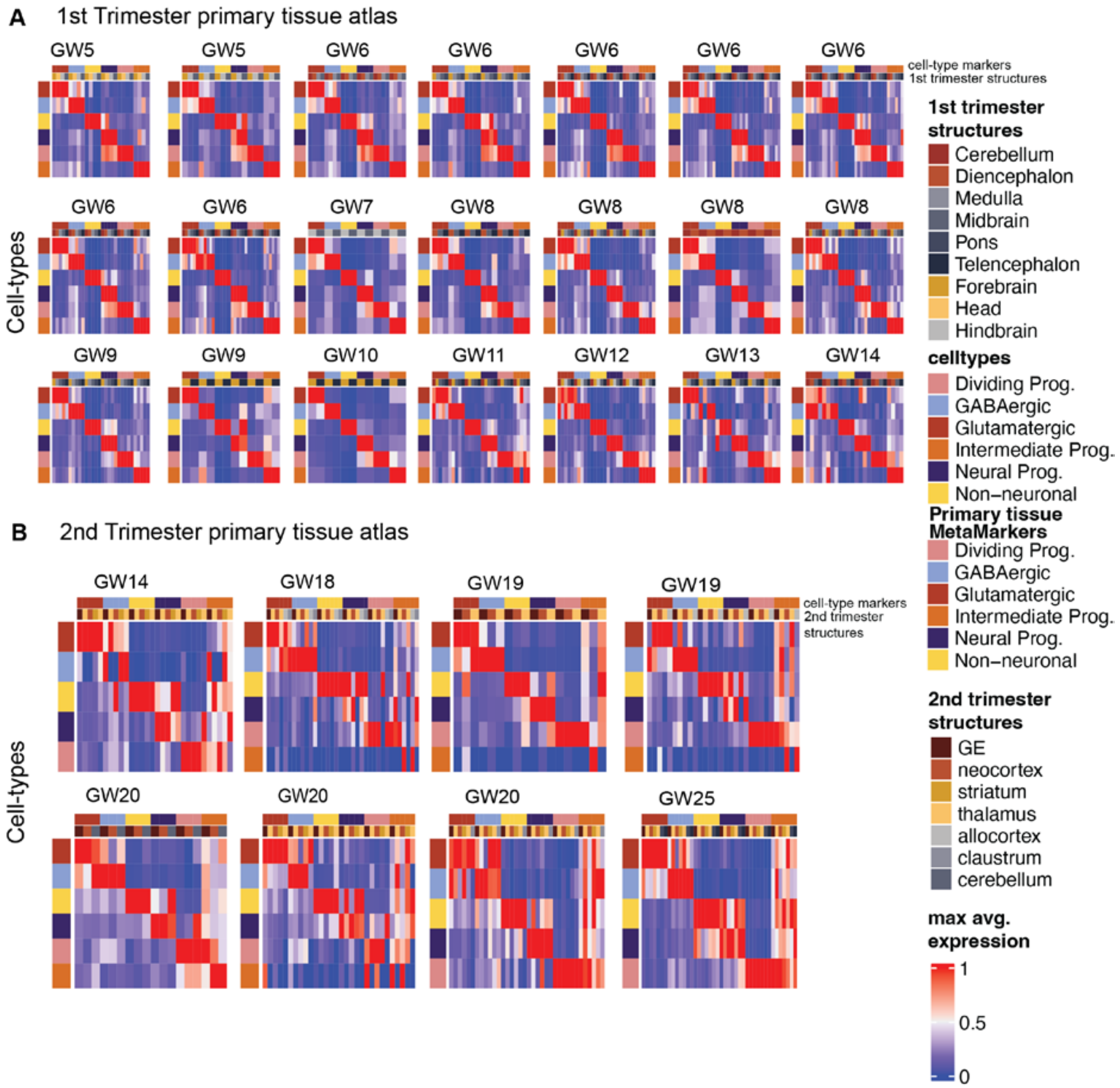


Figure S4.2. MetaMarkers as regionally robust primary tissue cell-type markers
A MetaMarkers exhibit cross-regional cell-type specificity. Heatmaps of maximum normalized average MetaMarker expression for cell-types and brain regions of the first trimester annotated primary tissue atlas. Cell-types comprise the rows with MetaMarker gene expression for cells from each annotated brain region comprising the columns. Data is maximum normalized per region/column.
B MetaMarkers exhibit cross-regional cell-type specificity. Heatmaps of maximum normalized average MetaMarker expression for cell-types and brain regions of the second trimester annotated primary tissue atlas. Cell-types comprise the rows with MetaMarker gene expression for cells from each annotated brain region comprising the columns. Data is maximum normalized per region/column.

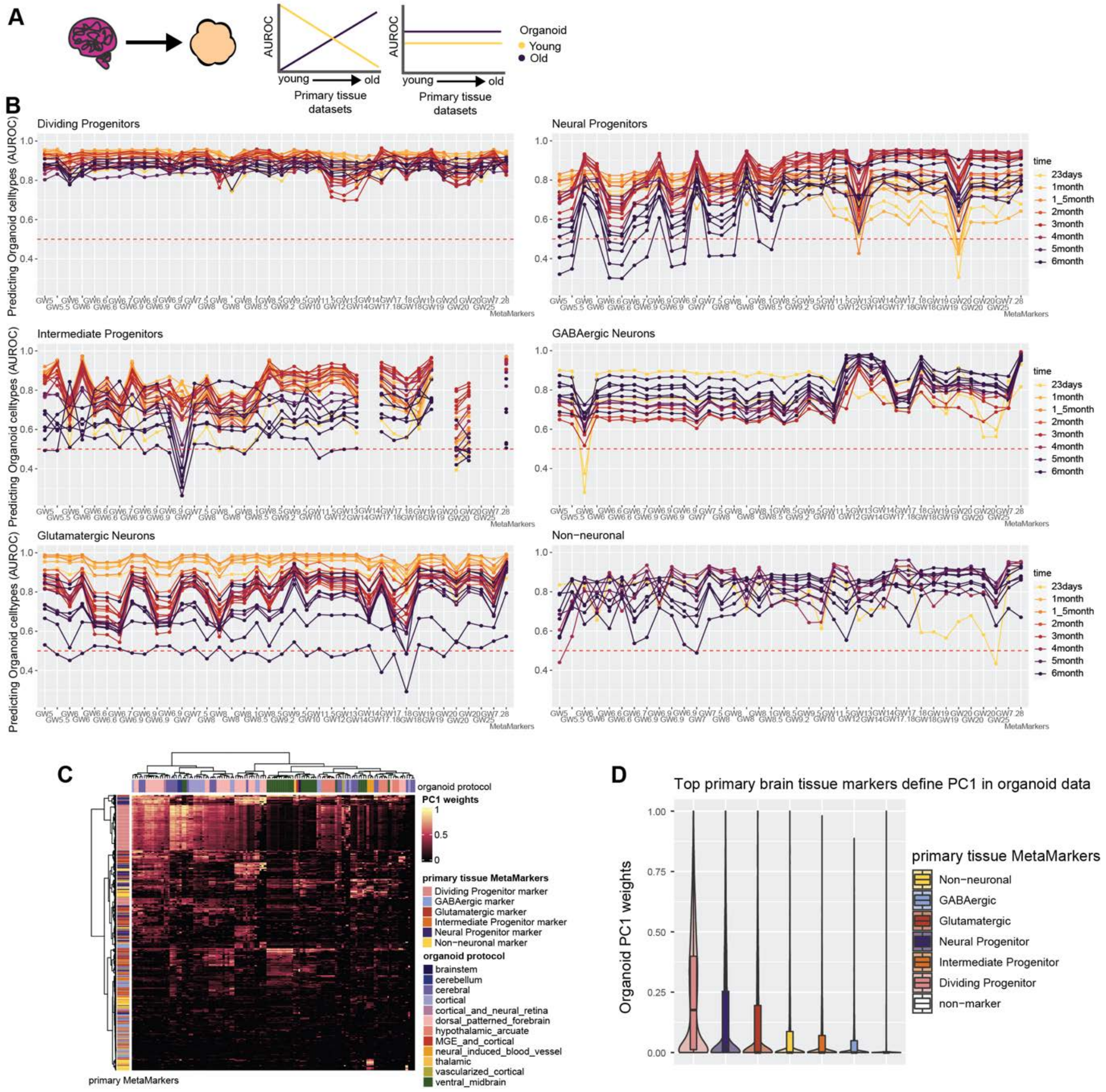


Figure S4.3. Primary tissue MetaMarkers consistently predict organoid cell-types across timepoints

A Schematic showing two potential outcomes when comparing cell-type marker expression between primary tissue and organoid data on a temporal axis. There may be a temporal relationship, with younger organoids recapitulating younger primary tissue marker expression over older primary tissue marker expression and vice versa for older organoids, or there may be no temporal relationship.

B Broad primary tissue cell-type markers have consistent performance predicting organoid annotations independent of temporal variation. Line plots showing the cell-type prediction AUROCs using top 100 markers from individual primary tissue datasets for all organoid time points. Primary tissue datasets on the x-axis are ordered from youngest to oldest.

C Primary tissue MetaMarkers define the first organoid principal component. Heatmap of normalized eigenvalues for primary tissue MetaMarkers within the first principal component of each organoid dataset.

D MetaMarker gene-set distributions of normalized PC1 eigenvalues across all organoid datasets.

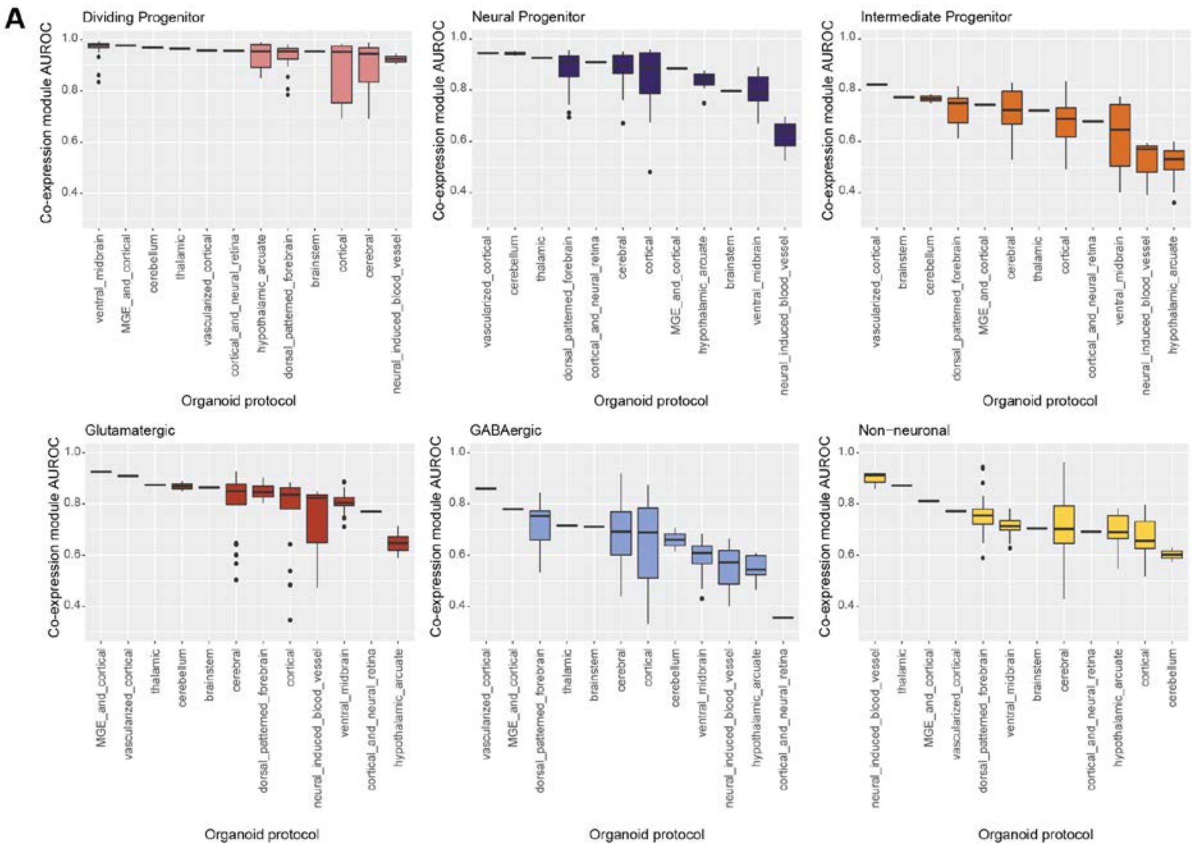


Figure S4.4. Intra-marker set MetaMarker co-expression varies over organoid protocols
A Organoids vary by protocol type for their primary tissue cell-type co-expression module scores. Boxplot distributions of co-expression module scores for the primary tissue MetaMarkers computed from organoid co-expression networks. Scores for organoid networks are grouped by organoid protocol type and ordered by their median score.

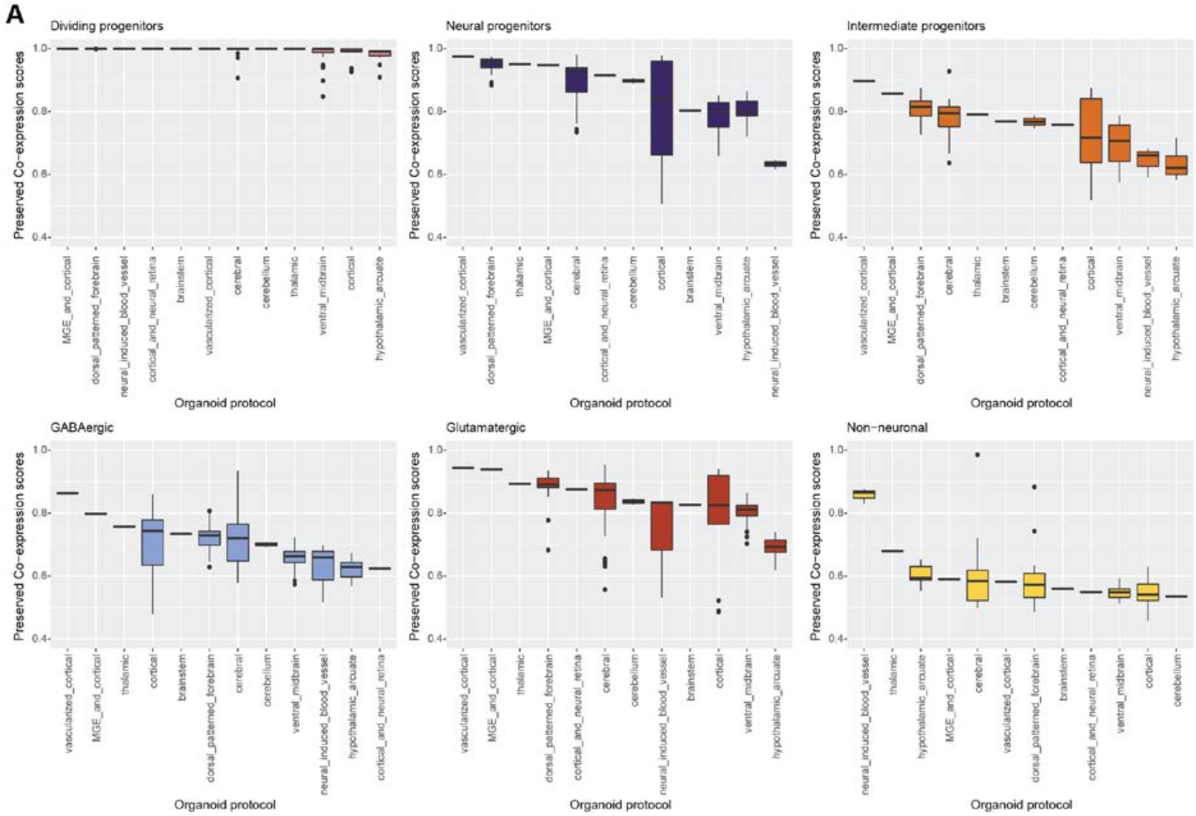


Figure S4.5. Preservation of MetaMarker set co-expression varies over organoid protocols
A Organoids vary by protocol type for their primary tissue cell-type preserved co-expression scores. Boxplot distributions of preserved co-expression scores for the primary tissue MetaMarkers computed from organoid co-expression networks. Scores for organoid networks are grouped by organoid protocol type and ordered by their median score.

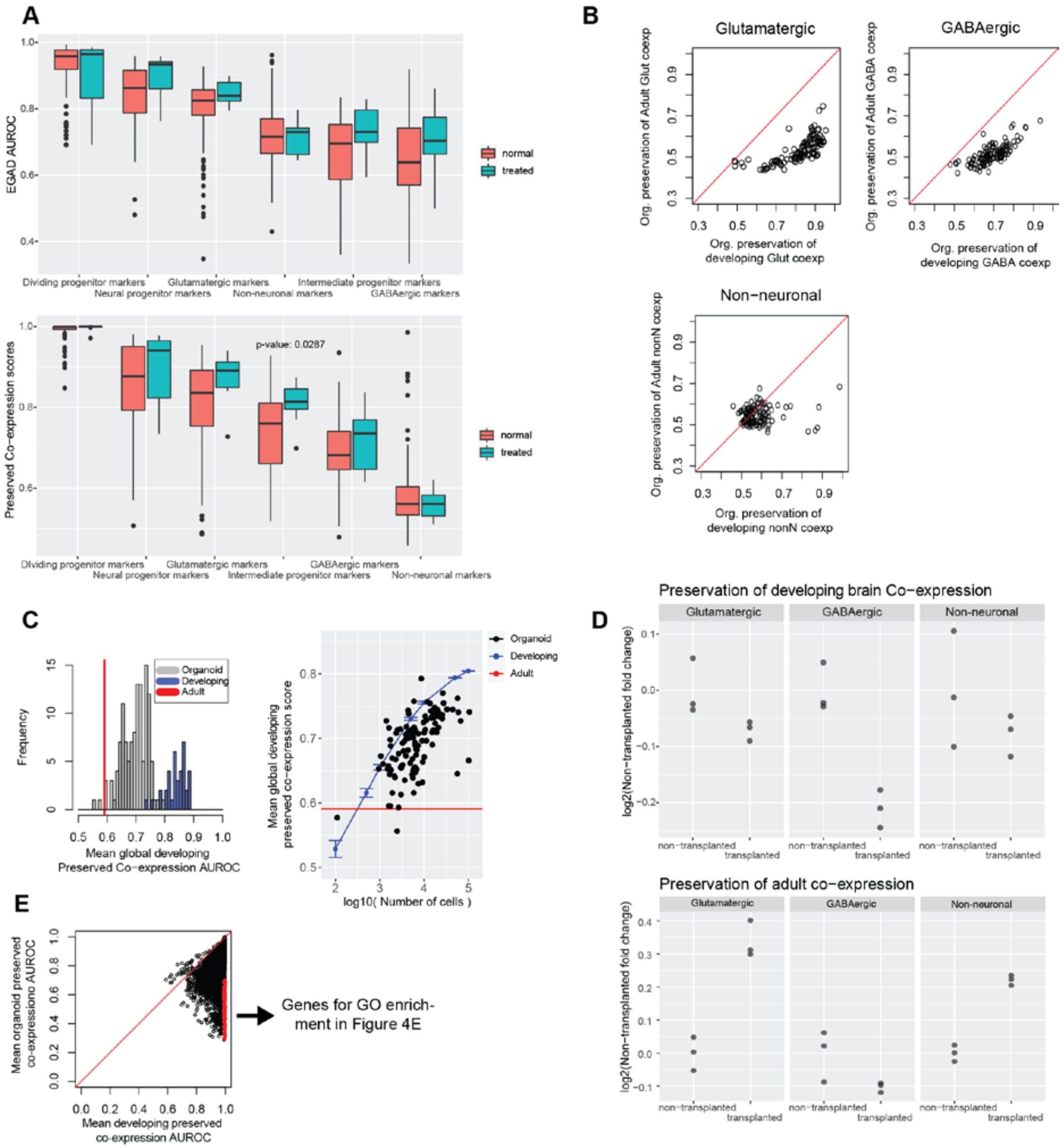

A
EGAD AUROC
1.0
0.8
0.6
0.4
Dividing progenitor markers
Neural progenitor markers
Glutamatergic markers
Non-neuronal markers
Intermediate progenitor markers
GABAergic markers
normal
treated
Preserved Co−expression scores
p-value: 0.0287
Dividing progenitor markers
Neural progenitor markers
Glutamatergic markers
Intermediate progenitor markers
GABAergic markers
Non-neuronal markers
B
Glutamatergic
Org. preservation of Adult Glut coexp
Org. preservation of developing Glut coexp
GABAergic
Org. preservation of Adult GABA coexp
Org. preservation of developing GABA coexp
Non−neuronal
Org. preservation of Adult nonN coexp
Org. preservation of developing nonN coexp
0.3
0.5
0.7
0.9
C
Frequency
Organoid
Developing
Adult
Mean global developing Preserved Co−expression AUROC
Mean global developing preserved co−expression score
log10(Number of cells)
D
Preservation of developing brain Co−expression
Glutamatergic
GABAergic
Non−neuronal
log2(Non−transplanted fold change)
non−transplanted
transplanted
Preservation of adult co−expression
E
Mean organoid preserved co-expressiono AUROC
Mean developing preserved co-expression AUROC
Genes for GO enrichment in Figure 4E

Figure S4.6. Neural organoids preserve co-expression of developing neural tissue over adult neural tissue

A Normal and treated organoids exhibit no differences in their recapitulation of primary tissue co-expression. Boxplots comparing either the co-expression module scores or preserved co-expression scores by cell-type between normal and treated organoids.

B Organoids preserve developing neuronal co-expression over adult co-expression. Scatterplots showing the preserved co-expression scores of either the top 100 developing brain MetaMarkers (x-axis) or the top 100 adult MetaMarkers (y-axis).

C Organoids lie between adult and developing brain data for global preservation of developing brain co-expression. Distributions of average preserved developing brain co-expression AUROCs across all genes for organoid and developing brain networks. The redline shows the performance of the adult co-expression network. The scatterplot plots the data in the histogram (y-axis) against the number of cells in each organoid dataset (x-axis). The blue line shows performance for a cell down-sampled developing brain dataset, with points representing the average performance over 10 random samples and the error bars showing ± 1 standard deviation.

D Transplanted organoids preserve adult co-expression over developing brain co-expression. Points represent the log2-fold change over the mean performance of the non-transplanted organoids for preserved co-expression scores.

E Organoids globally have low preserved developing brain co-expression of individual genes across the genome. Points show the average preserved developing brain co-expression AUROC of individual genes, comparing the average across developing brain networks (x-axis) against the average across organoid networks (y-axis). The points colored in red are genes with developing brain scores >= 0.99 and organoid scores < 0.70.

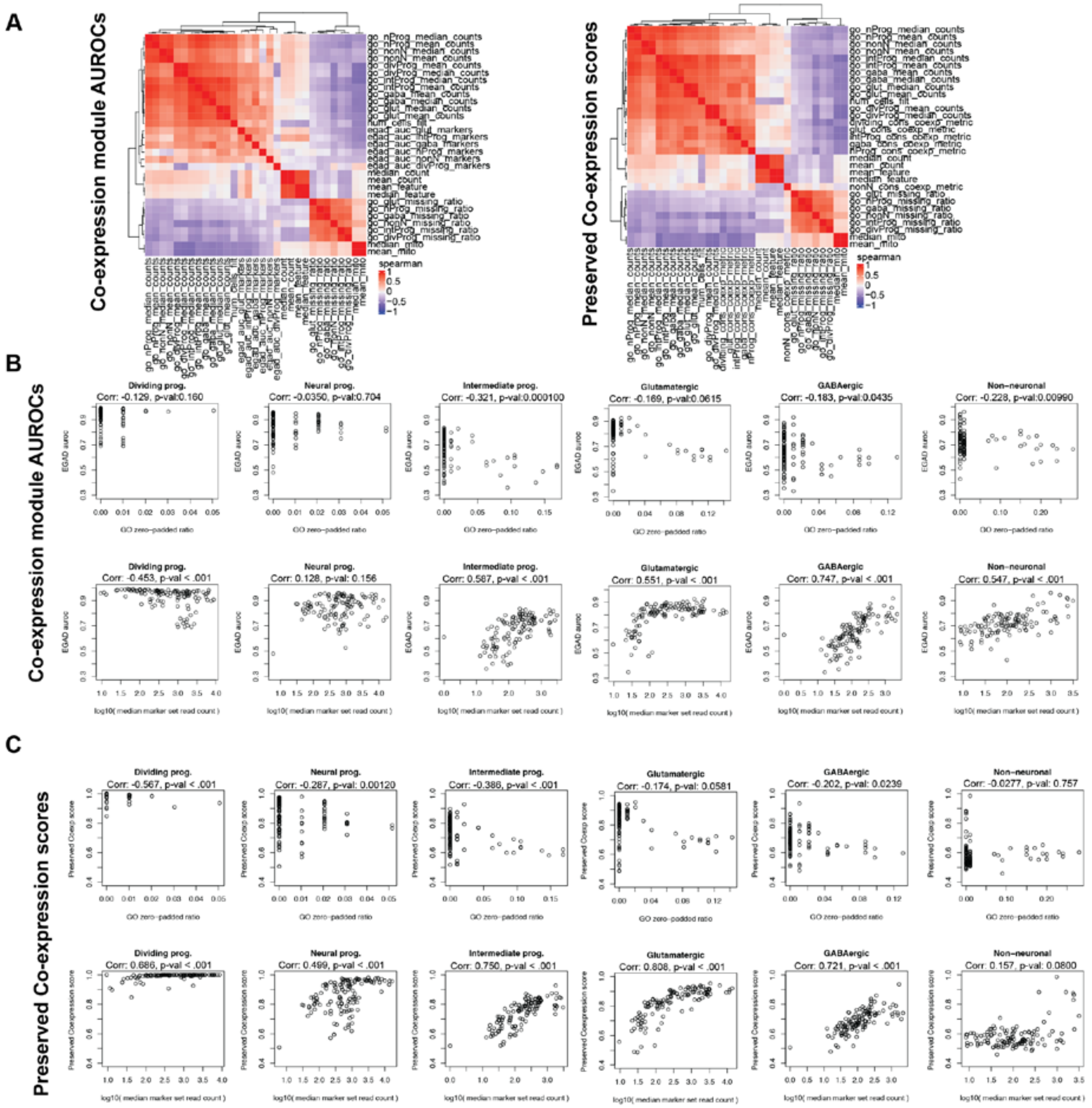


Figure S4.7. Strength of MetaMarker co-expression in organoids is related to expression levels
A Marker set expression and cell number are strongly correlated with co-expression performance across organoid datasets. Heatmaps of spearman correlations between either co-expression module scores or preserved co-expression scores and various technical features of each network/dataset, like marker set expression, dataset sequencing depth, number of cells in each dataset, and the zero-padding ratio of each marker set.
B Scatterplots of either the zero-padded ratio (top row) or marker set expression (bottom row) against the co-expression module scores for each cell-type across the organoid datasets.
C Scatterplots of either the zero-padded ratio (top row) or marker set expression (bottom row) against the preserved co-expression scores for each cell-type across the organoid datasets.

4.8. Chapter 4 summary:

In this work, we extract universal cell-type specific signatures from primary neural tissue data for use as a benchmark against highly heterogeneous neural organoid systems. We first identify genes that act as strong cell-type markers for primary neural tissue across extensive developmental timepoints (gestational weeks 5-25) and brain regions (15 regions), constituting universal primary tissue signatures. We then compare the co-expression relationships of these primary tissue markers in independent primary tissue data and across neural organoid systems, revealing that primary tissue samples recover consistently strong marker-set specific co-expression modules whereas organoids present highly variable marker set co-expression. We quantify the strength of preserved primary tissue co-expression in neural organoids for individual genes, marker sets, and genome-wide measures, demonstrating that while primary tissue has universally high performance, neural organoids range from zero to comparable primary tissue signal. In summary, this work presents a generalizable quantitative benchmark for grading the fidelity of neural lineage production within neural organoid models.

5. Conclusions and Perspectives

5.1. General summary

The work presented in this thesis investigates developmental lineage through several fronts: exploiting XCI as a marker for developmental lineage to characterize early lineage specification events during human development and 8 other non-human mammalian species, and employing meta-analytic approaches for benchmarking the primary tissue fidelity of *in vitro* neural differentiation. Broadly, this work revealed XCI is completed prior to any tissue differentiation in humans and the stochastically determined XCI ratio during embryogenesis is propagated through development to all tissues, with additional cell sampling events injecting lineage-specific XCI variability. We extended analysis of XCI variability to 8 other mammalian species, demonstrating models of embryonic stochasticity explain population XCI ratio variability consistently across species as opposed to genetic factors. Finally, we developed a quantitative benchmark for measuring the fidelity of neural organoid systems to primary neural tissues at gene, cell-type, and genome-wide scales applicable to a wide-range of heterogeneous neural organoid models. In summary, this work characterizes early developmental lineage events at an organismal scale across tissues in humans, puts forth a general model for explaining observed population-scale XCI ratio variability across mammalian species, and derives a generalizable benchmark for grading the successes and failures of current *in vitro* models for neural developmental lineages.

5.2. Discussion for Results chapter 2: cross-tissue variability in human XCI ratios

In this work, we exploited the random, permanent, and developmentally early nature of XCI to investigate characteristics of early lineage specification events during human development. By analyzing variance in XCI ratios across tissues and individuals, we showed human XCI is completed before tissue specification and the stochastically determined XCI ratio set during embryogenesis is a shared feature across all tissue lineages. We estimate a lower bound of 6-16 cells are fated for the embryonic epiblast lineage based on population-level variance in XCI ratios. Additionally, we provide lower bound estimates of the number of cells present during tissue-specific lineage specification for 46 different tissues. To conduct this analysis, we developed a method to estimate the ratio of XCI using unphased allele-specific expression, a highly scalable approach applicable to any bulk RNA-sequencing sample.

This work provides insight into the observed variance of XCI ratios in normal female populations, an area of ongoing debate (Brown and Robinson 2000; Migeon 1998; Clerc and Avner 2006; Peeters et al. 2016). Our results indicate that the initial embryonic XCI ratio is propagated through development and is a shared feature across all tissue lineages. This demonstrates the stochasticity of the initial choice for inactivation within the embryo has a measurable impact on XCI ratios in adult females. Importantly, GTEx donors presumably represent a phenotypically normal population; as such, our analysis captures XCI variance in the absence of potential drivers (X-linked diseases) of allelic-selection, representing the null distribution of XCI variation in adult females.

Additional contributors to the observed variance in XCI ratios across tissues may be genetic variation that can drive allelic selection over development (Brown and Robinson 2000; Schmidt and Sart 1992) or stochastic deviations in XCI ratios caused by developmental proliferation (Sun et al. 2021). In contrast to these models, we report strikingly consistent XCI ratios across tissues for individual donors, and, importantly, across tissues derived from different germ layers. If allelic-selection or stochastic deviations from proliferation were strong contributors to variance in XCI, we would not expect consistent XCI ratios across developmentally distant adult tissues. Nevertheless, it is unlikely that the initial embryonic

XCI ratio is propagated through development with perfect fidelity, which contextualizes our cell number estimates as lower bound estimates for the number of cells that must have been involved in XCI or lineage specification events. In general, our results suggest XCI ratios are broadly shared across tissues with lineage-specific stochasticity due to cell sampling effects during lineage-specification.

For the timing of XCI, there is a wealth of complimentary research on the exact molecular mechanisms (Dossin and Heard 2021; Vallot et al. 2017) that define the highly complex biological process of XCI. XCI is a continuous molecular process and recent studies from human embryos suggest the timing of XCI may overlap the lineage specification of the extraembryonic and embryonic tissues (Moreira de Mello et al. 2017; Petropoulos et al. 2016), which precedes germ layer specification. In this study, we aimed to interrogate timing of XCI as it relates to germ layer specification within the embryonic lineage. Any overlap in timing for the molecular process of XCI and extraembryonic/embryonic lineage specification will have no impact on our results and conclusions of shared variance in XCI within the embryonic lineage. The consideration of extraembryonic tissues provides the developmental context that XCI ratio variance within the germ layer lineages may be a combination of XCI stochasticity and cell sampling during embryonic epiblast specification.

One alternative model consistent with our results is the potential for rapid allelic changes in the time between XCI and germ layer specification, allowing for selection or drift to occur, with the XCI ratio then stabilized after germ layer specification. While possible, we find this improbable due to the small number of cell divisions estimated to occur between XCI and germ layer specification, as well as the lack of evidence for any continued effects after germ layer specification.

Our work is part of a broader history of using X-linked mosaicism as a useful tool for studying lineage relationships, with studies ranging from investigations of early lineage events in mice (Nesbitt 1971) to ascertaining tumor clonality (Linder and Gartler 1965). Typically, these approaches will capitalize on a single locus of the X-chromosome to determine XCI status (Boudewijns et al. 2007). One of our methodological contributions is demonstrating the allelic expression imbalance generated via XCI can be aggregated across multiple loci to provide near-perfect estimates of XCI ratios, even in the absence of phased information.

While the GTEx dataset aims to sample non-diseased tissues, we cannot rule out potential disease-states, genetic or otherwise, for all tissue samples, where disease may impact allelic selection and contribute to variance inn XCI ratios. When assessing escape from XCI, we focus on genes with constitutive rather than facultative signal and cannot make conclusions on likely tissue- or donor-specific escape. Our tissue-specific cell count estimations depend on the sample size of the given tissue and the number of tissues sampled for individual donors, both of which vary considerably across tissues and individuals. As such, these estimates are likely rough approximations that can be improved with additional tissue and donor sampling.

5.3. Discussion for Results chapter 3: cross-species population variability in XCI ratios

We utilized bulk RNA-seq samples to model tissue XCI ratios, establishing population-level distributions of XCI ratios across 9 mammalian species. Our analysis revealed substantial variation in XCI ratios among different mammalian populations, likely reflecting differences in the timing of XCI or embryonic/extra-embryonic lineage specification during development. We demonstrated models of embryonic stochasticity explain population-level XCI variability exceptionally well, providing estimates for the number of cells present during embryonic events. These cell count estimates represent either the number of cells present at the time of embryonic lineage specification or at the time of XCI, depending on the temporal ordering of extra-embryonic/embryonic lineage specification and XCI, which may vary across species. Furthermore, we examined the potential genetic correlates of XCI ratios and found a consistent lack of associations both at a broad level across the entire X-chromosome and for individual variants. This suggests that the inherent stochastic nature of XCI, rather than genetic factors, primarily drives population-level XCI variability across mammals.

The lack of cross-mammalian comparisons of population XCI variability has previously limited our understanding on the sources of XCI variability in mammals. The existence of XCE-alleles in laboratory mice(Cattanach and Isaacson 1965; Simmler et al. 1993; Sun et al. 2021; Calaway et al. 2013) has supported the hypothesis that a similar genetic mechanism can exist in humans and drive population XCI variability(Peeters et al. 2016), though evidence for XCE-alleles in human populations remains inconclusive(Bolduc et al. 2008) and data from other mammalian species is historically absent. Although genetic influences on XCI, particularly variants affecting XIST(Plenge et al. 1997) or disease-associated variants(Migeon 1971; Migeon et al. 1981; Devriendt et al. 1997; Plenge et al. 2002), have been identified, they do not constitute a general mechanism that can fully account for observed population-level XCI variability. Comprehensive assessment of genetic influence on XCI would require combined DNA and RNA sequencing data, which is challenging to perform at a large scale across mammalian populations. Our approach for extracting heterozygous variants from RNA-seq data(Werner et al. 2022), while providing a sample of genetic variability, is still able to assess hundreds of X-linked genes per species for associations with XCI and culminated in only weak evidence for limited genetic influence on XCI ratios. In contrast, we demonstrated models of embryonic stochasticity can explain population XCI variability with exceedingly small amounts of error consistently across mammalian species, providing a much more general explanation for population XCI variability.

Other potential contributors to XCI ratio variability other than those already discussed (X-linked disorders and XIST-variants) include genomic incompatibilities(Shorter et al. 2017) and stochastic allelic drift during development(Sun et al. 2021). We identified an association between the variance in X-linked allelic expression and the degree of inbreeding among several of the sampled species (Fig. 3.2B) as well as autosome-specific allelic imbalances in dog (Supp. Fig. 3.4). This suggests variability in X-linked allelic expression may be a combination of the sampled bulk XCI ratio along with broader genomic incompatibilities between the parental genomes(Shorter et al. 2017), dependent on the species. Our approach for excluding samples that exhibit global allelic imbalances (Supp. Fig. 3.4) is a powerful control that demonstrates the allelic-expression variability we sample from the X-chromosome is highly specific to XCI. Additionally, stochastic allelic drift through development(Sun et al. 2021) may potentially inject variability in XCI ratios outside of the initial random choice of allelic inactivation. While our previous cross-tissue analysis of XCI ratios in humans(Werner et al. 2022) revealed consistency in XCI ratios across developmentally distant tissues, suggesting allelic drift is not a strong influencing factor in XCI ratio variability, similar cross-tissue data for non-human mammals is lacking. Overall, these potential contributing factors to XCI variability contextualize our cell count estimates as lower bound estimates for the number of cells required to produce the observed XCI ratio variability as explained purely through embryonic stochasticity.

We revealed population variability in XCI ratios itself is conserved across adult mammalian populations, raising the question as to why stochasticity in XCI evolved in the first place. An alternative route for achieving dosage compensation of the X-chromosome is the non-random inactivation of a specific allele, as evidenced by imprinted inactivation of the paternal X-allele in marsupial species(Deakin et al. 2009) and in the extra-embryonic lineages of rodents(Takagi and Sasaki 1975; Wake et al. 1976). One putative explanation for both the random inactivation in mammals and imprinted inactivation in marsupials is a general lack of selective pressure for the expression of either parental allele; both X-alleles are largely identical in terms of fitness. It is well known the X-chromosome is depleted of genetic variability compared to the autosomes(Sachidanandam et al. 2001), suggesting the X-chromosome is under higher rates of evolution (models of increased positive or purifying selection are widely debated(Payseur et al. 2002; Avery 1984; Casto et al. 2010; Veeramah et al. 2014)). These routes for genetic homogeneity of the X-chromosome could favor the evolution of either imprinted or random inactivation in the face of selective pressure to achieve dosage compensation.

As a general model, the depletion of X-linked genetic variability also explains the lack of evidence we observe for broad allelic-selection or individual variants driving XCI ratio variability in mammalian populations, as both parental alleles are principally equivalent. This is of course not the case in the presence of disease variants, but X-linked disease genetics cannot explain the pervasive XCI ratio variability we report across mammalian species. While we are able to identify genes associated with increased XCI ratios that have prior evidence for contributing to highly skewed XCI in disease cases, the effect sizes and population frequencies of these variants are small in our sample populations. In conclusion, the general lack of X-linked genetic variability positions the inherent stochasticity of XCI during embryogenesis as the basis for the observed XCI ratio variability in mammalian populations.

5.4. Discussion for Results chapter 4: meta-analysis of preserved co-expression between primary neural tissue and neural organoids

Through the use of meta-analytic differential expression and co-expression, we are able to provide cell-type specific measurements of human neural organoids' current capacity to replicate primary tissue biology. We extracted broad cell-type markers that define primary brain tissue cell-types across a large temporal axis (GW5 – 25) and across numerous heterogenous brain regions to act as a generalizable primary tissue reference for organoids that also vary temporally and regionally (by protocol). By quantifying intra-marker set co-expression and the preservation of co-expression across networks, we revealed human neural organoids lie on a spectrum of near-zero to near-identical recapitulation of primary tissue cell-type specific co-expression in comparison to primary tissue data. We made our aggregate primary tissue reference data and methods for measuring preserved co-expression publicly available as an R package to aid in the quality control and protocol development of future human neural organoids.

Prior work comparing primary brain tissue and neural organoid systems demonstrated organoids can produce cell-types(Velasco et al. 2019; Bhaduri et al. 2020) and morphological structures(Sakaguchi et al. 2015; Qian et al. 2020) similar to primary tissues and are capable of modeling temporal(Gordon et al. 2021; Uzquiano et al. 2022; Amiri et al. 2018) and regional(Lancaster et al. 2013; Bhaduri et al. 2020; Qian et al. 2016; Xiang et al. 2017) primary tissue variation. Multiple lines of evidence support these findings such as assessments of cytoarchitecture and cell-type proportions(Lancaster et al. 2013; Velasco et al. 2019; Tanaka et al. 2020; Agboola et al. 2021), whole transcriptome and marker gene expression correlations(Camp et al. 2015; Bhaduri et al. 2020), and comparisons of co-expression modules(Pollen et al. 2019; Gordon et al. 2021; Cheroni et al. 2022; Luo et al. 2016). Our meta-analytic approach is able to quantify these field-wide observations within a generalizable framework, recapitulating that organoids model broad primary tissue biology with our specific approach offering several key advancements for primary tissue/organoid comparisons. First, we derive quantifications of preserved primary tissue co-expression that can be extended from individual genes to the entire genome and, second, we place organoid co-expression in reference to robust meta-analytic primary tissue performance providing a general benchmark for protocol development and quality control across heterogeneous organoid systems.

A key aspect of our study design is our cross-validation of primary tissue differential expression and co-expression. We demonstrated that temporally and regionally heterogenous primary tissue data are able to strongly recapitulate our meta-analytic primary tissue marker gene expression and co-expression relationships. This meta-analytic primary tissue performance defines a clear benchmark for gauging the fidelity of organoid models, where organoids that produce functionally equivalent primary tissue cell-types are expected to perform comparably to primary tissue data. In our assessment across 12 different organoid differentiation protocols, we showed a subset of protocols produce organoids with comparable cell-type specific co-expression to primary tissue data, demonstrating high primary tissue fidelity is possible with current methods. While we employ a broad approach sampling across temporal and regional variation to optimize for generalizability, more precisely matched primary tissue data for specific organoid timepoints or protocols is better suited for comparisons studying more subtle variation.

Certainly, while comparisons between primary tissue and organoid systems at a high-resolution of cell-type annotation are of interest, our results centered on broad cell-types at the cell-class level constitute a critical foundation for these more fine-tuned investigations of organoids. Cell-type specification within the brain involves complex spatial and temporal mechanisms(Nowakowski et al. 2017) to produce the high cellular heterogeneity we observe, with the exact resolution of meaningful cell-type annotations still being actively debated and posing a general conceptual challenge within the field of single-cell genomics(Zeng 2022). We focus here on establishing methods for assessing consistent and accurate production of primary tissue cell-types at the class-level within organoids as a critical actionable first step towards increasing primary tissue fidelity across variable organoid differentiation protocols, with meta-analysis at a higher resolution of cell-type annotations (e.g., MGE and CGE interneurons, layer-specific excitatory neurons, progenitor subtypes) as an exciting future venture once class-level fidelity in organoids is consistently achieved.

One exciting application for the use of neural organoid systems is the study of a wide-range of human neurological diseases using human *in vitro* models(Chen et al. 2019; Eichmüller and Knoblich 2022), which critically depends on the *in vivo* fidelity of cell-types produced in organoids. Neural organoids have been used to model and investigate human disorders of neurodevelopmental (Lancaster et al. 2013; Mariani et al. 2015), neuropsychiatric(Notaras et al. 2022; Stachowiak et al. 2017; Dixon and Muotri 2023), and neurodegenerative(Smits et al. 2019; Chen et al. 2021; Szebényi et al. 2021) nature, as well as infectious diseases(Qian et al. 2016; Garcez et al. 2016; Pellegrini et al. 2020). It is essential that organoid systems model *in vivo* cell-types with extreme fidelity to fully realize the therapeutic potential of human organoids and ensure findings in these *in vitro* models are not specific to potential artifactual or inaccurate *in vitro* biology. While our results demonstrate that high primary tissue fidelity in organoids is currently methodologically possible, we also report a high degree of variability across organoids and studies/protocols indicating a remaining methodological gap. The broad applicability of our meta-analytic approach offers the potential for benchmarking primary tissue fidelity across numerous organoid protocols, aiding in increasing the quality of neural organoids for use in a wide-range of human health-related translational investigations.

The generalizable and flexible nature of our analysis is well suited to aid in the development of organoid differentiation protocols and the general quality control of neural organoids. Our results demonstrate the type of experiments possible through comparing preserved co-expression across organoid experimental variables, such as the differences in preserved co-expression between organoids grown in vertical or orbital shakers, as well as between transplanted or non-transplanted organoids. Importantly, our broad sampling across organoid protocols enabled clear identification of promising avenues for increasing organoid primary tissue fidelity. The strong performance across cell-types for the vascularized protocol we assessed suggests vascularized protocols as a route forward for global increases in primary fidelity. Additionally, our findings of specific ECM-related genes with consistent poorly preserved primary tissue co-expression in organoids suggests investigations into the interactions between Matrigel or other ECM-substrates and organoids may lead to general protocol adjustments for increasing primary tissue fidelity(Kozlowski et al. 2021). Looking beyond neural organoids, our framework for quantifying preserved co-expression can be applied to other organoid systems granted there is sufficient annotated primary tissue data to act as a reference.

5.5. Future directions

A large component of this thesis explores the utility in assessments of allelic-imbalances for the study of developmental lineage, an approach that can be extended beyond the X-chromosome and within organoid systems for more comprehensive and targeted investigations of developmental lineage.

Autosomal allelic imbalances for studying causal lineage events

The allelic imbalance of the X-chromosome via XCI is useful in studying developmental lineage because it occurs early and affects an entire chromosome, but most importantly, it is permanent and inherited down lineages. The early timing and broad chromosomal effect of XCI produces a strong, robust signal of allele-specific expression that is easy to detect, situating XCI as a natural starting point for investigating developmental lineages through allelic imbalances. Importantly, any inherited allele-specific effect can also be utilized in the study of developmental lineage; the X-chromosome simply being the one with the largest effect size. The methods established in this thesis for detecting and modeling X-chromosomal allelic imbalances are equally applicable to autosomal loci and can be used to investigate the interplay between allelic imbalances and developmental lineage on a genome-wide scale.

The observation of autosomal random monoallelic expression (RME) in various contexts (tissue-specific and/or cell-type specific) is long-standing, yet its functional relevance remains unclear in most cases (allele-specific expression of olfactory receptors as one notable exception). One hypothesis is RME is a mechanism for modulating expression levels of genes in a lineage-specific manner. If a particular gene dosage is required for a given cell fate, the epigenetic regulation of allele-specific expression, or RME in other words, is one route that canalizes gene dosage within a lineage. In the same way that variability in X-chromosome allelic-expression is informative for characteristics surrounding the time of XCI and ordering lineage events, variability in autosomal allelic-expression can also be informative for lineage specification events, with the addition of identifying genes that were causal for imparting functionality within a lineage. Taking a similar cross-tissue and cross-species approach as the XCI work presented in this thesis, expanding analysis to autosomal allelic imbalances stands to provide a genome-wide interrogation of lineage specification events at multiple scales.

Combining organoid systems with the study of allelic imbalances for investigating lineage specification events

The power of using autosomal imbalances for investigating lineage specification is fully realized when sampling over known distinct lineages, such as cross-tissue comparisons or, in a more controlled manner, organoids. Autosomal loci with variability in allelic-expression specific to an individual lineage are likely instances of epigenetic regulation required for that specific lineage. While enriching for specific lineages to sample over is difficult when working with *in vivo* tissues, *in vitro* organoids offer the opportunity for the scalable production of specific lineages. For example, with neural organoids there are numerous protocols to generate excitatory cortical organoids as well as inhibitory subpallial organoids, representing the lineage split between excitatory and inhibitory neurons. Identifying allelic imbalances specific to these lineages will be informative for identifying likely causal epigenetic regulation in the formation of excitatory or inhibitory neurons. A similar experimental design can be used over any set of lineages, with organoid systems offering a high degree of experimental control over the production of specific lineages.

Organoids that fail to produce comparable lineages to *in vivo* tissues are also potentially informative for identifying causal genes in lineage production. The generalizable benchmark for measuring primary tissue fidelity of neural organoids presented in this thesis will be particularly useful for such approaches. As one example, out of the variety of differentiation protocols claiming to produce excitatory neurons within neural organoids, our quantification of preserved co-expression demonstrates not all organoid excitatory neurons are equal to primary neural tissues. Exploring epigenetic landscapes via allelic imbalances between organoids with and without successful generation of excitatory neurons stands to again identify likely causal epigenetic regulation required for proper excitatory neuron production.

6. References

Agboola OS, Hu X, Shan Z, Wu Y, Lei L. 2021. Brain organoid: a 3D technology for investigating cellular composition and interactions in human neurological development and disease models in vitro. *Stem Cell Research & Therapy* **12**: 430.

Allen RC, Zoghbi HY, Moseley AB, Rosenblatt HM, Belmont JW. 1992. Methylation of HpaII and HhaI sites near the polymorphic CAG repeat in the human androgen-receptor gene correlates with X chromosome inactivation. *Am J Hum Genet* **51**: 1229–1239.

Amiri A, Coppola G, Scuderi S, Wu F, Roychowdhury T, Liu F, Pochareddy S, Shin Y, Safi A, Song L, et al. 2018. Transcriptome and epigenome landscape of human cortical development modeled in brain organoids. *Science* **362**: eaat6720.

Amos-Landgraf JM, Cottle A, Plenge RM, Friez M, Schwartz CE, Longshore J, Willard HF. 2006. X Chromosome–Inactivation Patterns of 1,005 Phenotypically Unaffected Females. *Am J Hum Genet* **79**: 493–499.

Andersen J, Revah O, Miura Y, Thom N, Amin ND, Kelley KW, Singh M, Chen X, Thete MV, Walczak EM, et al. 2020. Generation of Functional Human 3D Cortico-Motor Assembloids. *Cell* **183**: 1913-1929.e26.

Avery PJ. 1984. The population genetics of haplo-diploids and X-linked genes. *Genetics Research* **44**: 321–341.

Balaton BP, Fornes O, Wasserman WW, Brown CJ. 2021. Cross-species examination of X-chromosome inactivation highlights domains of escape from silencing. *Epigenetics Chromatin* **14**: 12.

Ballouz S, Weber M, Pavlidis P, Gillis J. 2017. EGAD: ultra-fast functional analysis of gene networks. *Bioinformatics* **33**: 612–614.

Barr ML, Bertram EG. 1949. A Morphological Distinction between Neurones of the Male and Female, and the Behaviour of the Nucleolar Satellite during Accelerated Nucleoprotein Synthesis. *Nature* **163**: 676–677.

Barr ML, Carr DH. 1960. Sex Chromatin, Sex Chromosomes and Sex Anomalies. *Can Med Assoc J* **83**: 979–986.

Belmont JW. 1996. Genetic control of X inactivation and processes leading to X-inactivation skewing. *Am J Hum Genet* **58**: 1101–1108.

Benito-Kwiecinski S, Giandomenico SL, Sutcliffe M, Riis ES, Freire-Pritchett P, Kelava I, Wunderlich S, Martin U, Wray GA, McDole K, et al. 2021. An early cell shape transition drives evolutionary expansion of the human forebrain. *Cell* **184**: 2084-2102.e19.

Berletch JB, Ma W, Yang F, Shendure J, Noble WS, Disteche CM, Deng X. 2015. Escape from X Inactivation Varies in Mouse Tissues. *PLoS Genet* **11**: e1005079.

Bhaduri A, Andrews MG, Mancia Leon W, Jung D, Shin D, Allen D, Jung D, Schmunk G, Haeussler M, Salma J, et al. 2020. Cell stress in cortical organoids impairs molecular subtype specification. *Nature* **578**: 142–148.

Birey F, Andersen J, Makinson CD, Islam S, Wei W, Huber N, Fan HC, Metzler KRC, Panagiotakos G, Thom N, et al. 2017. Assembly of functionally integrated human forebrain spheroids. *Nature* **545**: 54–59.

Bittel DC, Theodoro MF, Kibiryeva N, Fischer W, Talebizadeh Z, Butler MG. 2008. Comparison of X-chromosome inactivation patterns in multiple tissues from human females. *J Med Genet* **45**: 309–313.

Bolduc V, Chagnon P, Provost S, Dubé M-P, Belisle C, Gingras M, Mollica L, Busque L. 2008. No evidence that skewing of X chromosome inactivation patterns is transmitted to offspring in humans. https://www.jci.org/articles/view/33166/pdf (Accessed November 13, 2020).

Bonora G, Disteche CM. 2017. Structural aspects of the inactive X chromosome. *Philosophical Transactions of the Royal Society B: Biological Sciences* **372**: 20160357.

Boudewijns M, van Dongen JJM, Langerak AW. 2007. The Human Androgen Receptor X-Chromosome Inactivation Assay for Clonality Diagnostics of Natural Killer Cell Proliferations. *J Mol Diagn* **9**: 337–344.

Brown C, Robinson W. 2000. The causes and consequences of random and non-random X chromosome inactivation in humans: X chromosome inactivation in humans. *Clinical Genetics* **58**: 353–363.

Brown CJ, Hendrich BD, Rupert JL, Lafrenière RG, Xing Y, Lawrence J, Willard HF. 1992. The human XIST gene: Analysis of a 17 kb inactive X-specific RNA that contains conserved repeats and is highly localized within the nucleus. *Cell* **71**: 527–542.

Calaway JD, Lenarcic AB, Didion JP, Wang JR, Searle JB, McMillan L, Valdar W, Villena FP-M de. 2013. Genetic Architecture of Skewed X Inactivation in the Laboratory Mouse. *PLOS Genetics* **9**: e1003853.

Camp JG, Badsha F, Florio M, Kanton S, Gerber T, Wilsch-Bräuninger M, Lewitus E, Sykes A, Hevers W, Lancaster M, et al. 2015. Human cerebral organoids recapitulate gene expression programs of fetal neocortex development. *Proc Natl Acad Sci USA* **112**: 15672–15677.

Carlson M. 2019. org.Hs.eg.db: Genome wide annotation for Human.

Carrel L, Willard HF. 2005. X-inactivation profile reveals extensive variability in X-linked gene expression in females. *Nature* **434**: 400–404.

Casto AM, Li JZ, Absher D, Myers R, Ramachandran S, Feldman MW. 2010. Characterization of X-Linked SNP genotypic variation in globally distributed human populations. *Genome Biology* **11**: R10.

Cattanach BM, Isaacson JH. 1965. Genetic control over the inactivation of autosomal genes attached to the X-chromosome. *Z Vererbungsl* **96**: 313–323.

Chen HI, Song H, Ming G. 2019. Applications of Human Brain Organoids to Clinical Problems. *Developmental Dynamics* **248**: 53–64.

Chen X, Sun G, Tian E, Zhang M, Davtyan H, Beach TG, Reiman EM, Blurton-Jones M, Holtzman DM, Shi Y. 2021. Modeling Sporadic Alzheimer's Disease in Human Brain Organoids under Serum Exposure. *Adv Sci (Weinh)* **8**: 2101462.

Cheroni C, Trattaro S, Caporale N, López-Tobón A, Tenderini E, Sebastiani S, Troglio F, Gabriele M, Bressan RB, Pollard SM, et al. 2022. Benchmarking brain organoid recapitulation of fetal corticogenesis. *Transl Psychiatry* **12**: 1–16.

Clerc P, Avner P. 2006. Random X-chromosome inactivation: skewing lessons for mice and men. *Current Opinion in Genetics & Development* **16**: 246–253.

Corrò C, Novellasdemunt L, Li VSW. 2020. A brief history of organoids. *American Journal of Physiology-Cell Physiology* **319**: C151–C165.

Crow M, Paul A, Ballouz S, Huang ZJ, Gillis J. 2016. Exploiting single-cell expression to characterize co-expression replicability. *Genome Biology* **17**: 101.

Deakin JE, Chaumeil J, Hore TA, Marshall Graves JA. 2009. Unravelling the evolutionary origins of X chromosome inactivation in mammals: insights from marsupials and monotremes. *Chromosome Res* **17**: 671–685.

Devriendt K, Matthijs G, Legius E, Schollen E, Blockmans D, van Geet C, Degreef H, Cassiman JJ, Fryns JP. 1997. Skewed X-chromosome inactivation in female carriers of dyskeratosis congenita. *Am J Hum Genet* **60**: 581–587.

Dixon TA, Muotri AR. 2023. Advancing preclinical models of psychiatric disorders with human brain organoid cultures. *Mol Psychiatry* **28**: 83–95.

Dixon-McDougall T, Brown CJ. 2022. Multiple distinct domains of human XIST are required to coordinate gene silencing and subsequent heterochromatin formation. *Epigenetics & Chromatin* **15**: 6.

Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, Gingeras TR. 2013. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**: 15–21.

Dossin F, Heard E. 2021. The Molecular and Nuclear Dynamics of X-Chromosome Inactivation. *Cold Spring Harb Perspect Biol* a040196.

Durinck S, Spellman PT, Birney E, Huber W. 2009. Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt. *Nat Protoc* **4**: 1184–1191.

Eichmüller OL, Knoblich JA. 2022. Human cerebral organoids — a new tool for clinical neurology research. *Nat Rev Neurol* **18**: 661–680.

Eiraku M, Watanabe K, Matsuo-Takasaki M, Kawada M, Yonemura S, Matsumura M, Wataya T, Nishiyama A, Muguruma K, Sasai Y. 2008. Self-Organized Formation of Polarized Cortical Tissues from ESCs and Its Active Manipulation by Extrinsic Signals. *Cell Stem Cell* **3**: 519–532.

Eraslan G, Drokhlyansky E, Anand S, Fiskin E, Subramanian A, Slyper M, Wang J, Van Wittenberghe N, Rouhana JM, Waldman J, et al. 2022. Single-nucleus cross-tissue molecular reference maps toward understanding disease gene function. *Science* **376**: eabl4290.

Eura N, Matsui TK, Luginbühl J, Matsubayashi M, Nanaura H, Shiota T, Kinugawa K, Iguchi N, Kiriyama T, Zheng C, et al. 2020. Brainstem Organoids From Human Pluripotent Stem Cells. *Frontiers in Neuroscience* **14**. https://www.frontiersin.org/articles/10.3389/fnins.2020.00538 (Accessed January 26, 2023).

Evans HJ, Ford CE, Lyon MF, Gray J. 1965. DNA Replication and Genetic Expression in Female Mice with Morphologically Distinguishable X Chromosomes. *Nature* **206**: 900–903.

Fair SR, Julian D, Hartlaub AM, Pusuluri ST, Malik G, Summerfied TL, Zhao G, Hester AB, Ackerman WE, Hollingsworth EW, et al. 2020. Electrophysiological Maturation of Cerebral Organoids Correlates with Dynamic Morphological and Cellular Development. *Stem Cell Reports* **15**: 855–868.

Fang H, Deng X, Disteche CM. 2021. X-factors in human disease: impact of gene content and dosage regulation. *Human Molecular Genetics* **30**: R285–R295.

Fang H, Disteche CM, Berletch JB. 2019. X Inactivation and Escape: Epigenetic and Structural Features. *Front Cell Dev Biol* **7**. https://www.frontiersin.org/articles/10.3389/fcell.2019.00219/full (Accessed December 18, 2019).

Feng W, Schriever H, Jiang S, Bais A, Wu H, Kostka D, Li G. 2022. Computational profiling of hiPSC-derived heart organoids reveals chamber defects associated with NKX2-5 deficiency. *Commun Biol* **5**: 1–18.

Fialkow PJ. 1973. Primordial cell pool size and lineage relationships of five human cell types*. *Annals of Human Genetics* **37**: 39–48.

Fialkow PJ, Gartler SM, Yoshida A. 1967. Clonal origin of chronic myelocytic leukemia in man. *Proc Natl Acad Sci U S A* **58**: 1468–1471.

Fialkow PJ, Sagebiel RW, Gartler SM, Rimoin DL. 1971. Multiple cell origin of hereditary neurofibromas. *N Engl J Med* **284**: 298–300.

Fischer S, Gillis J. 2021. How many markers are needed to robustly determine a cell's type? *iScience* **24**: 103292.

Fleck JS, Jansen SMJ, Wollny D, Zenk F, Seimiya M, Jain A, Okamoto R, Santel M, He Z, Camp JG, et al. 2022. Inferring and perturbing cell fate regulomes in human brain organoids. *Nature* 1–8.

Frankish A, Diekhans M, Jungreis I, Lagarde J, Loveland JE, Mudge JM, Sisu C, Wright JC, Armstrong J, Barnes I, et al. 2021. GENCODE 2021. *Nucleic Acids Research* **49**: D916–D923.

Gandini E, Gartler SM. 1969. Glucose-6-phosphate Dehydrogenase Mosaicism for studying the Development of Blood Cell Precursors. *Nature* **224**: 599–600.

Gandini E, Gartler SM, Angioni G, Argiolas N, Dell'Acqua G. 1968. Developmental implications of multiple tissue studies in glucose-6-phosphate dehydrogenase-deficient heterozygotes. *Proc Natl Acad Sci USA* **61**: 945–948.

Garcez PP, Loiola EC, Madeiro da Costa R, Higa LM, Trindade P, Delvecchio R, Nascimento JM, Brindeiro R, Tanuri A, Rehen SK. 2016. Zika virus impairs growth in human neurospheres and brain organoids. *Science* **352**: 816–818.

Gart JJ. 1970. A Locally Most Powerful Test for the Symmetric Folded Binomial Distribution. *Biometrics* **26**: 129–138.

Gartler SM, Gandini E, Angioni G, Argiolas N. 1969. Glucose-6 phosphate dehydrogenase mosaicism: utilization as a tracer in the study of the development of hair root cells*. *Annals of Human Genetics* **33**: 171–176.

Gartler SM, Ziprkowski L, Krakowski A, Ezra R, Szeinberg A, Adam A. 1966. Glucose-6-phosphate dehydrogenase mosaicism as a tracer in the study of hereditary multiple trichoepithelioma. *Am J Hum Genet* **18**: 282–287.

Geens M, Chuva De Sousa Lopes SM. 2017. X chromosome inactivation in human pluripotent stem cells as a model for human development: back to the drawing board? *Hum Reprod Update* **23**: 520–532.

Gel B, Serra E. 2017. karyoploteR : an R / Bioconductor package to plot customizable genomes displaying arbitrary data. *Bioinformatics* **33**: 3088–3090.

Ghimire S, Mantziou V, Moris N, Martinez Arias A. 2021. Human gastrulation: The embryo and its models. *Developmental Biology* **474**: 100–108.

Gordon A, Yoon S-J, Tran SS, Makinson CD, Park JY, Andersen J, Valencia AM, Horvath S, Xiao X, Huguenard JR, et al. 2021. Long-term maturation of human cortical organoids matches key early postnatal transitions. *Nat Neurosci* **24**: 331–342.

Gu Z, Eils R, Schlesner M. 2016. Complex heatmaps reveal patterns and correlations in multidimensional genomic data. *Bioinformatics*.

Hagen SH, Henseling F, Hennesen J, Savel H, Delahaye S, Richert L, Ziegler SM, Altfeld M. 2020. Heterogeneous Escape from X Chromosome Inactivation Results in Sex Differences in Type I IFN Responses at the Single Human pDC Level. *Cell Rep* **33**: 108485.

Heard E, Disteche CM. 2006. Dosage compensation in mammals: fine-tuning the expression of the X chromosome. *Genes Dev* **20**: 1848–1867.

Hoon B de, Monkhorst K, Riegman P, Laven JSE, Gribnau J. 2015. Buccal swab as a reliable predictor for X inactivation ratio in inaccessible tissues. *Journal of Medical Genetics* **52**: 784–790.

Huang W-K, Wong SZH, Pather SR, Nguyen PTT, Zhang F, Zhang DY, Zhang Z, Lu L, Fang W, Chen L, et al. 2021. Generation of hypothalamic arcuate organoids from human induced pluripotent stem cells. *Cell Stem Cell* **28**: 1657-1670.e10.

Huch M, Dorrell C, Boj SF, van Es JH, Li VSW, van de Wetering M, Sato T, Hamer K, Sasaki N, Finegold MJ, et al. 2013. In vitro expansion of single Lgr5+ liver stem cells induced by Wnt-driven regeneration. *Nature* **494**: 247–250.

Jorstad NL, Song JHT, Exposito-Alonso D, Suresh H, Castro N, Krienen FM, Yanny AM, Close J, Gelfand E, Travaglini KJ, et al. 2022. *Comparative transcriptomics reveals human-specific cortical features*. Neuroscience http://biorxiv.org/lookup/doi/10.1101/2022.09.19.508480 (Accessed January 11, 2023).

Kanton S, Boyle MJ, He Z, Santel M, Weigert A, Sanchís-Calleja F, Guijarro P, Sidow L, Fleck JS, Han D, et al. 2019. Organoid single-cell genomic atlas uncovers human-specific features of brain development. *Nature* **574**: 418–422.

Kim HJ, O'Hara-Wright M, Kim D, Loi TH, Lim BY, Jamieson RV, Gonzalez-Cordero A, Yang P. 2023. Comprehensive characterization of fetal and mature retinal cell identity to assess the fidelity of retinal organoids. *Stem Cell Reports* **18**: 175–189.

Kloska A, Jakóbkiewicz-Banecka J, Tylki-Szymańska A, Czartoryska B, Węgrzyn G. 2011. Female Hunter syndrome caused by a single mutation and familial XCI skewing: implications for other X-linked disorders. *Clin Genet* **80**: 459–465.

Knudsen GPS, Neilson TCS, Pedersen J, Kerr A, Schwartz M, Hulten M, Bailey MES, Ørstavik KH. 2006. Increased skewing of X chromosome inactivation in Rett syndrome patients and their mothers. *Eur J Hum Genet* **14**: 1189–1194.

Kozlowski MT, Crook CJ, Ku HT. 2021. Towards organoid culture without Matrigel. *Commun Biol* **4**: 1–15.

Lancaster MA, Knoblich JA. 2014. Generation of cerebral organoids from human pluripotent stem cells. *Nat Protoc* **9**: 2329–2340.

Lancaster MA, Renner M, Martin C-A, Wenzel D, Bicknell LS, Hurles ME, Homfray T, Penninger JM, Jackson AP, Knoblich JA. 2013. Cerebral organoids model human brain development and microcephaly. *Nature* **501**: 373–379.

Larsson AJM, Coucoravas C, Sandberg R, Reinius B. 2019. X-chromosome upregulation is driven by increased burst frequency. *Nat Struct Mol Biol* **26**: 963–969.

Lee J, Shah M, Ballouz S, Crow M, Gillis J. 2020. CoCoCoNet: conserved and comparative co-expression across a diverse set of species. *Nucleic Acids Research* **48**: W566–W571.

Leek JT, Scharpf RB, Bravo HC, Simcha D, Langmead B, Johnson WE, Geman D, Baggerly K, Irizarry RA. 2010. Tackling the widespread and critical impact of batch effects in high-throughput data. *Nat Rev Genet* **11**: 733–739.

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, 1000 Genome Project Data Processing Subgroup. 2009. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**: 2078–2079.

Linder D, Gartler SM. 1965. Glucose-6-Phosphate Dehydrogenase Mosaicism: Utilization as a Cell Marker in the Study of Leiomyomas. *Science* **150**: 67–69.

Lonsdale J, Thomas J, Salvatore M, Phillips R, Lo E, Shad S, Hasz R, Walters G, Garcia F, Young N, et al. 2013. The Genotype-Tissue Expression (GTEx) project. *Nat Genet* **45**: 580–585.

Luo C, Lancaster MA, Castanon R, Nery JR, Knoblich JA, Ecker JR. 2016. Cerebral Organoids Recapitulate Epigenomic Signatures of the Human Fetal Brain. *Cell Reports* **17**: 3369–3384.

Lyon MF. 1961. Gene Action in the X -chromosome of the Mouse (Mus musculus L.). *Nature* **190**: 372–373.

Lyon MF. 1972. X-chromosome inactivation and developmental patterns in mammals. *Biol Rev Camb Philos Soc* **47**: 1–35.

Magaraki A, Loda A, Gontan C, Merzouk S, Sleddens-Linkels E, Meek S, Baarends WM, Burdon T, Gribnau J. 2019. A novel approach to differentiate rat embryonic stem cells in vitro reveals a role for RNF12 in activation of X chromosome inactivation. *Sci Rep* **9**: 6068.

Mariani J, Coppola G, Zhang P, Abyzov A, Provini L, Tomasini L, Amenduni M, Szekely A, Palejev D, Wilson M, et al. 2015. FOXG1-Dependent Dysregulation of GABA/Glutamate Neuron Differentiation in Autism Spectrum Disorders. *Cell* **162**: 375–390.

Mayhew CN, Singhania R. 2023. A review of protocols for brain organoids and applications for disease modeling. *STAR Protocols* **4**: 101860.

McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M, et al. 2010. The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res* **20**: 1297–1303.

Mclaren A. 1972. Biological Sciences: Numerology of Development. *Nature* **239**: 274–276.

McMahon A, Fosten M, Monk M. 1983. X-chromosome inactivation mosaicism in the three germ layers and the germ line of the mouse embryo. *J Embryol Exp Morphol* **74**: 207–220.

Mead BE, Ordovas-Montanes J, Braun AP, Levy LE, Bhargava P, Szucs MJ, Ammendolia DA, MacMullan MA, Yin X, Hughes TK, et al. 2018. Harnessing single-cell genomics to improve the physiological fidelity of organoid-derived cell types. *BMC Biology* **16**: 62.

Migeon B. 2013. *Females Are Mosaics: X Inactivation and Sex Differences in Disease*. Oxford University Press https://oxfordmedicine.com/view/10.1093/med/9780199927531.001.0001/med-9780199927531 (Accessed July 28, 2021).

Migeon BR. 2016. An overview of X inactivation based on species differences. *Semin Cell Dev Biol* **56**: 111–116.

Migeon BR. 1998. Non-random X chromosome inactivation in mammalian cells. *Cytogenet Cell Genet* **80**: 142–148.

Migeon BR. 1971. Studies of skin fibroblasts from 10 families with HGPRT deficiency, with reference in X-chromosomal inactivation. *Am J Hum Genet* **23**: 199–210.

Migeon BR, Moser HW, Moser AB, Axelman J, Sillence D, Norum RA. 1981. Adrenoleukodystrophy: evidence for X linkage, inactivation, and selection favoring the mutant allele in heterozygous cells. *Proc Natl Acad Sci U S A* **78**: 5066–5070.

Miller JM, Malenfant RM, David P, Davis CS, Poissant J, Hogg JT, Festa-Bianchet M, Coltman DW. 2014. Estimating genome-wide heterozygosity: effects of demographic history and marker type. *Heredity (Edinb)* **112**: 240–247.

Miura Y, Li M-Y, Birey F, Ikeda K, Revah O, Thete MV, Park J-Y, Puno A, Lee SH, Porteus MH, et al. 2020. Generation of human striatal organoids and cortico-striatal assembloids from human pluripotent stem cells. *Nat Biotechnol* **38**: 1421–1430.

Monteiro J, Derom C, Vlietinck R, Kohn N, Lesser M, Gregersen PK. 1998. Commitment to X Inactivation Precedes the Twinning Event in Monochorionic MZ Twins. *The American Journal of Human Genetics* **63**: 339–346.

Morcillo P, Qin Y, Peña G, Mosenthal AC, Livingston DH, Spolarics Z. 2020. Directional X Chromosome Skewing of White Blood Cells from Subjects with Heterozygous Mosaicism for the Variant IRAK1 Haplotype. *Inflammation* **43**: 370–381.

Moreira de Mello JC, Fernandes GR, Vibranovski MD, Pereira LV. 2017. Early X chromosome inactivation during human preimplantation development revealed by single-cell RNA-sequencing. *Sci Rep* **7**: 10794.

Muguruma K, Nishiyama A, Kawakami H, Hashimoto K, Sasai Y. 2015. Self-Organization of Polarized Cerebellar Tissue in 3D Culture of Human Pluripotent Stem Cells. *Cell Reports* **10**: 537–550.

Mutzel V, Okamoto I, Dunkel I, Saitou M, Giorgetti L, Heard E, Schulz EG. 2019. A symmetric toggle switch explains the onset of random X inactivation in different mammals. *Nat Struct Mol Biol* **26**: 350–360.

Nascimento JM, Saia-Cereda VM, Sartore RC, da Costa RM, Schitine CS, Freitas HR, Murgu M, de Melo Reis RA, Rehen SK, Martins-de-Souza D. 2019. Human Cerebral Organoids and Fetal Brain Tissue Share Proteomic Similarities. *Front Cell Dev Biol* **7**: 303.

Naumova AK, Plenge RM, Bird LM, Leppert M, Morgan K, Willard HF, Sapienza C. 1996. Heritability of X chromosome--inactivation phenotype in a large family. *Am J Hum Genet* **58**: 1111–1119.

Nayler S, Agarwal D, Curion F, Bowden R, Becker EBE. 2021. High-resolution transcriptional landscape of xeno-free human induced pluripotent stem cell-derived cerebellar organoids. *Sci Rep* **11**: 12959.

Nesbitt MN. 1971. X chromosome inactivation mosaicism in the mouse. *Developmental Biology* **26**: 252–263.

Newman AM, Steen CB, Liu CL, Gentles AJ, Chaudhuri AA, Scherer F, Khodadoust MS, Esfahani MS, Luca BA, Steiner D, et al. 2019. Determining cell type abundance and expression from bulk tissues with digital cytometry. *Nat Biotechnol* **37**: 773–782.

Notaras M, Lodhi A, Dündar F, Collier P, Sayles NM, Tilgner H, Greening D, Colak D. 2022. Schizophrenia is defined by cell-specific neuropathology and multiple neurodevelopmental mechanisms in patient-derived cerebral organoids. *Mol Psychiatry* **27**: 1416–1434.

Nowakowski TJ, Bhaduri A, Pollen AA, Alvarado B, Mostajo-Radji MA, Di Lullo E, Haeussler M, Sandoval-Espinosa C, Liu SJ, Velmeshev D, et al. 2017. Spatiotemporal gene expression trajectories reveal developmental hierarchies of the human cortex. *Science* **358**: 1318–1323.

Ohno S. 1966. *Sex Chromosomes and Sex Linked Genes*. Springer Berlin, Heidelberg.

Okamoto I, Nakamura T, Sasaki K, Yabuta Y, Iwatani C, Tsuchiya H, Nakamura S, Ema M, Yamamoto T, Saitou M. 2021. The X chromosome dosage compensation program during the development of cynomolgus monkeys. *Science* **374**: eabd8887.

Okamoto I, Patrat C, Thépot D, Peynot N, Fauque P, Daniel N, Diabangouaya P, Wolf J-P, Renard J-P, Duranthon V, et al. 2011. Eutherian mammals use diverse strategies to initiate X-chromosome inactivation during development. *Nature* **472**: 370–374.

O'Leary NA, Wright MW, Brister JR, Ciufo S, Haddad D, McVeigh R, Rajput B, Robbertse B, Smith-White B, Ako-Adjei D, et al. 2016. Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res* **44**: D733-745.

Pagès H, Carlson M, Falcon S, Li N. 2022. AnnotationDbi: Manipulation of SQLite-based annotations in Bioconductor.

Payseur BA, Cutter AD, Nachman MW. 2002. Searching for Evidence of Positive Selection in the Human Genome Using Patterns of Microsatellite Variability. *Molecular Biology and Evolution* **19**: 1143–1153.

Peeters SB, Yang C, Brown CJ. 2016. Have humans lost control: The elusive X-controlling element. *Seminars in Cell & Developmental Biology* **56**: 71–77.

Pellegrini L, Albecka A, Mallery DL, Kellner MJ, Paul D, Carter AP, James LC, Lancaster MA. 2020. SARS-CoV-2 Infects the Brain Choroid Plexus and Disrupts the Blood-CSF Barrier in Human Brain Organoids. *Cell Stem Cell* **27**: 951-961.e5.

Petropoulos S, Edsgärd D, Reinius B, Deng Q, Panula SP, Codeluppi S, Plaza Reyes A, Linnarsson S, Sandberg R, Lanner F. 2016. Single-Cell RNA-Seq Reveals Lineage and X Chromosome Dynamics in Human Preimplantation Embryos. *Cell* **165**: 1012–1026.

Plenge RM, Hendrich BD, Schwartz C, Arena JF, Naumova A, Sapienza C, Winter RM, Willard HF. 1997. A promoter mutation in the XIST gene in two unrelated families with skewed X-chromosome inactivation. *Nat Genet* **17**: 353–356.

Plenge RM, Stevenson RA, Lubs HA, Schwartz CE, Willard HF. 2002. Skewed X-chromosome inactivation is a common feature of X-linked mental retardation disorders. *Am J Hum Genet* **71**: 168–173.

Pollen AA, Bhaduri A, Andrews MG, Nowakowski TJ, Meyerson OS, Mostajo-Radji MA, Di Lullo E, Alvarado B, Bedolli M, Dougherty ML, et al. 2019. Establishing Cerebral Organoids as Models of Human-Specific Brain Evolution. *Cell* **176**: 743-756.e17.

Posynick BJ, Brown CJ. 2019. Escape From X-Chromosome Inactivation: An Evolutionary Perspective. *Front Cell Dev Biol* **7**. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6817483/ (Accessed December 18, 2019).

Qian X, Nguyen HN, Song MM, Hadiono C, Ogden SC, Hammack C, Yao B, Hamersky GR, Jacob F, Zhong C, et al. 2016. Brain-Region-Specific Organoids Using Mini-bioreactors for Modeling ZIKV Exposure. *Cell* **165**: 1238–1254.

Qian X, Su Y, Adam CD, Deutschmann AU, Pather SR, Goldberg EM, Su K, Li S, Lu L, Jacob F, et al. 2020. Sliced Human Cortical Organoids for Modeling Distinct Cortical Layer Formation. *Cell Stem Cell* **26**: 766-781.e9.

Quinlan AR, Hall IM. 2010. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**: 841–842.

R Core Team. 2021. R: A language and environment for statistical computing. *R Foundation for Statistical Computing, Vienna, Austria*.

R Core Team. 2023. R: A Language and Environment for Statistical Computing.

Ramos-Ibeas P, Sang F, Zhu Q, Tang WWC, Withey S, Klisch D, Wood L, Loose M, Surani MA, Alberio R. 2019. Pluripotency and X chromosome dynamics revealed in pig pre-gastrulating embryos by single cell analysis. *Nat Commun* **10**: 500.

Revah O, Gore F, Kelley KW, Andersen J, Sakai N, Chen X, Li M-Y, Birey F, Yang X, Saw NL, et al. 2022. Maturation and circuit integration of transplanted human cortical organoids. *Nature* **610**: 319–326.

Robertson SP, Twigg SRF, Sutherland-Smith AJ, Biancalana V, Gorlin RJ, Horn D, Kenwrick SJ, Kim CA, Morava E, Newbury-Ecob R, et al. 2003. Localized mutations in the gene encoding the cytoskeletal protein filamin A cause diverse malformations in humans. *Nat Genet* **33**: 487–491.

Robinson JT, Thorvaldsdóttir H, Winckler W, Guttman M, Lander ES, Getz G, Mesirov JP. 2011. Integrative genomics viewer. *Nat Biotechnol* **29**: 24–26.

Rozowsky J, Drenkow J, Yang YT, Gursoy G, Galeev T, Borsari B, Epstein CB, Xiong K, Xu J, Gao J, et al. 2021. Multi-tissue integrative analysis of personal epigenomes. *bioRxiv* 2021.04.26.441442.

Rozowsky J, Gao J, Borsari B, Yang YT, Galeev T, Gürsoy G, Epstein CB, Xiong K, Xu J, Li T, et al. 2023. The EN-TEx resource of multi-tissue personal epigenomes & variant-impact models. *Cell* **186**: 1493-1511.e40.

Sachidanandam R, Weissman D, Schmidt SC, Kakol JM, Stein LD, Marth G, Sherry S, Mullikin JC, Mortimore BJ, Willey DL, et al. 2001. A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. *Nature* **409**: 928–933.

Sachs N, Papaspyropoulos A, Zomer-van Ommen DD, Heo I, Böttinger L, Klay D, Weeber F, Huelsz-Prince G, Iakobachvili N, Amatngalim GD, et al. 2019. Long-term expanding human airway organoids for disease modeling. *The EMBO Journal* **38**: e100300.

Sakaguchi H, Kadoshima T, Soen M, Narii N, Ishida Y, Ohgushi M, Takahashi J, Eiraku M, Sasai Y. 2015. Generation of functional hippocampal neurons from self-organizing human embryonic stem cell-derived dorsomedial telencephalic tissue. *Nat Commun* **6**: 8896.

Sarieva K, Mayer S. 2021. The Effects of Environmental Adversities on Human Neocortical Neurogenesis Modeled in Brain Organoids. *Frontiers in Molecular Biosciences* **8**. https://www.frontiersin.org/articles/10.3389/fmolb.2021.686410 (Accessed February 1, 2023).

Sato T, Vries RG, Snippert HJ, van de Wetering M, Barker N, Stange DE, van Es JH, Abo A, Kujala P, Peters PJ, et al. 2009. Single Lgr5 stem cells build crypt-villus structures in vitro without a mesenchymal niche. *Nature* **459**: 262–265.

Schmidt M, Sart DD. 1992. Functional disomies of the X chromosome influence the cell selection and hence the X inactivation pattern in females with balanced X-autosome translocations: A review of 122 cases. *American Journal of Medical Genetics* **42**: 161–169.

Sherry ST, Ward M-H, Kholodov M, Baker J, Phan L, Smigielski EM, Sirotkin K. 2001. dbSNP: the NCBI database of genetic variation. *Nucleic Acids Research* **29**: 308–311.

Shorter JR, Odet F, Aylor DL, Pan W, Kao C-Y, Fu C-P, Morgan AP, Greenstein S, Bell TA, Stevans AM, et al. 2017. Male Infertility Is Responsible for Nearly Half of the Extinction Observed in the Mouse Collaborative Cross. *Genetics* **206**: 557–572.

Shvetsova E, Sofronova A, Monajemi R, Gagalova K, Draisma HHM, White SJ, Santen GWE, Chuva de Sousa Lopes SM, Heijmans BT, van Meurs J, et al. 2019. Skewed X-inactivation is common in the general female population. *Eur J Hum Genet* **27**: 455–465.

Simmler MC, Cattanach BM, Rasberry C, Rougeulle C, Avner P. 1993. Mapping the murine Xce locus with (CA)n repeats. *Mamm Genome* **4**: 523–530.

Skinnider MA, Squair JW, Foster LJ. 2019. Evaluating measures of association for single-cell transcriptomics. *Nat Methods* **16**: 381–386.

Smits LM, Reinhardt L, Reinhardt P, Glatza M, Monzel AS, Stanslowsky N, Rosato-Siri MD, Zanon A, Antony PM, Bellmann J, et al. 2019. Modeling Parkinson's disease in midbrain-like organoids. *npj Parkinsons Dis* **5**: 1–8.

Sozzi E, Nilsson F, Kajtez J, Parmar M, Fiorenzano A. 2022. Generation of Human Ventral Midbrain Organoids Derived from Pluripotent Stem Cells. *Current Protocols* **2**: e555.

Stachowiak EK, Benson CA, Narla ST, Dimitri A, Chuye LEB, Dhiman S, Harikrishnan K, Elahi S, Freedman D, Brennand KJ, et al. 2017. Cerebral organoids reveal early cortical maldevelopment in schizophrenia—computational anatomy and genomics, role of FGFR1. *Transl Psychiatry* **7**: 1–24.

Stuart JM, Segal E, Koller D, Kim SK. 2003. A Gene-Coexpression Network for Global Discovery of Conserved Genetic Modules. *Science* **302**: 249–255.

Sun KY, Oreper D, Schoenrock SA, McMullan R, Giusti-Rodríguez P, Zhabotynsky V, Miller DR, Tarantino LM, Pardo-Manuel de Villena F, Valdar W. 2021. Bayesian modeling of skewed X inactivation in genetically diverse mice identifies a novel Xce allele associated with copy number changes. *Genetics* **218**: iyab034.

Suong DNA, Imamura K, Inoue I, Kabai R, Sakamoto S, Okumura T, Kato Y, Kondo T, Yada Y, Klein WL, et al. 2021. Induction of inverted morphology in brain organoids by vertical-mixing bioreactors. *Commun Biol* **4**: 1–13.

Szebényi K, Wenger LMD, Sun Y, Dunn AWE, Limegrover CA, Gibbons GM, Conci E, Paulsen O, Mierau SB, Balmus G, et al. 2021. Human ALS/FTD brain organoid slice cultures display distinct early astrocyte and targetable neuronal pathology. *Nat Neurosci* **24**: 1542–1554.

Szelinger S, Malenica I, Corneveaux JJ, Siniard AL, Kurdoglu AA, Ramsey KM, Schrauwen I, Trent JM, Narayanan V, Huentelman MJ, et al. 2014. Characterization of X Chromosome Inactivation Using Integrated Analysis of Whole-Exome and mRNA Sequencing. *PLOS ONE* **9**: e113036.

Takagi N, Sasaki M. 1975. Preferential inactivation of the paternally derived X chromosome in the extraembryonic membranes of the mouse. *Nature* **256**: 640–642.

Takasato M, Er PX, Chiu HS, Maier B, Baillie GJ, Ferguson C, Parton RG, Wolvetang EJ, Roost MS, Lopes SMC de S, et al. 2015. Kidney organoids from human iPS cells contain multiple lineages and model human nephrogenesis. *Nature* **526**: 564–568.

Tanaka Y, Cakir B, Xiang Y, Sullivan GJ, Park I-H. 2020. Synthetic Analyses of Single-Cell Transcriptomes from Multiple Brain Organoids and Fetal Brain. *Cell Reports* **30**: 1682-1689.e3.

Tukiainen T, Villani A-C, Yen A, Rivas MA, Marshall JL, Satija R, Aguirre M, Gauthier L, Fleharty M, Kirby A, et al. 2017. Landscape of X chromosome inactivation across human tissues. *Nature* **550**: 244–248.

Urbakh VYu. 1967. Statistical Testing of Differences in Causal Behaviour of Two Morphologically Indistinguishable Objects. *Biometrics* **23**: 137–143.

Uzquiano A, Kedaigle AJ, Pigoni M, Paulsen B, Adiconis X, Kim K, Faits T, Nagaraja S, Antón-Bolaños N, Gerhardinger C, et al. 2022. Proper acquisition of cell class identity in organoids allows definition of fate specification programs of the human cerebral cortex. *Cell* **185**: 3770-3788.e27.

Vallot C, Patrat C, Collier AJ, Huret C, Casanova M, Liyakat Ali TM, Tosolini M, Frydman N, Heard E, Rugg-Gunn PJ, et al. 2017. XACT Noncoding RNA Competes with XIST in the Control of X Chromosome Activity during Human Early Development. *Cell Stem Cell* **20**: 102–111.

van de Geijn B, McVicker G, Gilad Y, Pritchard JK. 2015. WASP: allele-specific software for robust molecular quantitative trait locus discovery. *Nat Methods* **12**: 1061–1063.

van den Berg IM, Laven JSE, Stevens M, Jonkers I, Galjaard R-J, Gribnau J, Hikke van Doorninck J. 2009. X Chromosome Inactivation Is Initiated in Human Preimplantation Embryos. *Am J Hum Genet* **84**: 771–779.

Vavrek MJ. 2020. fossil: Palaeoecological and Palaeogeographical Analysis Tools. https://CRAN.R-project.org/package=fossil (Accessed February 6, 2023).

Veeramah KR, Gutenkunst RN, Woerner AE, Watkins JC, Hammer MF. 2014. Evidence for Increased Levels of Positive and Negative Selection on the X Chromosome versus Autosomes in Humans. *Molecular Biology and Evolution* **31**: 2267–2282.

Velasco S, Kedaigle AJ, Simmons SK, Nash A, Rocha M, Quadrato G, Paulsen B, Nguyen L, Adiconis X, Regev A, et al. 2019. Individual brain organoids reproducibly form cell diversity of the human cerebral cortex. *Nature* **570**: 523–527.

Waddington CH. 1942a. Canalization of Development and the Inheritance of Acquired Characters. *Nature* **150**: 563–565.

Waddington CH. 1942b. The epigenotype. *Int J Epidemiol* **41**: 10–13.

Wake N, Takagi N, Sasaki M. 1976. Non-random inactivation of X chromosome in the rat yolk sac. *Nature* **262**: 580–581.

Watanabe K, Kamiya D, Nishiyama A, Katayama T, Nozaki S, Kawasaki H, Watanabe Y, Mizuseki K, Sasai Y. 2005. Directed differentiation of telencephalic precursors from embryonic stem cells. *Nat Neurosci* **8**: 288–296.

Werner JM, Ballouz S, Hover J, Gillis J. 2022. Variability of cross-tissue X-chromosome inactivation characterizes timing of human embryonic lineage specification events. *Developmental Cell* **57**: 1995-2008.e5.

Wickham H. 2016. ggplot2: Elegant Graphics for Data Analysis. *Springer-Verlag New York*. https://ggplot2.tidyverse.org.

Winham SJ, Larson NB, Armasu SM, Fogarty ZC, Larson MC, McCauley BM, Wang C, Lawrenson K, Gayther S, Cunningham JM, et al. 2019. Molecular signatures of X chromosome inactivation and associations with clinical outcomes in epithelial ovarian cancer. *Hum Mol Genet* **28**: 1331–1342.

Wu H, Luo J, Yu H, Rattner A, Mo A, Wang Y, Smallwood PM, Erlanger B, Wheelan SJ, Nathans J. 2014. Cellular Resolution Maps of X Chromosome Inactivation: Implications for Neural Development, Function, and Disease. *Neuron* **81**: 103–119.

Xiang Y, Tanaka Y, Cakir B, Patterson B, Kim K-Y, Sun P, Kang Y-J, Zhong M, Liu X, Patra P, et al. 2019. hESC-Derived Thalamic Organoids Form Reciprocal Projections When Fused with Cortical Organoids. *Cell Stem Cell* **24**: 487-497.e7.

Xiang Y, Tanaka Y, Patterson B, Kang Y-J, Govindaiah G, Roselaar N, Cakir B, Kim K-Y, Lombroso AP, Hwang S-M, et al. 2017. Fusion of Regionally Specified hPSC-Derived Organoids Models Human Brain Development and Interneuron Migration. *Cell Stem Cell* **21**: 383-398.e7.

Yu B, van Tol HTA, Stout TAE, Roelen BAJ. 2020. Initiation of X Chromosome Inactivation during Bovine Embryo Development. *Cells* **9**: 1016.

Zeng H. 2022. What is a cell type and how to define it? *Cell* **185**: 2739–2755.

Zhang B, Horvath S. 2005. A General Framework for Weighted Gene Co-Expression Network Analysis. *Statistical Applications in Genetics and Molecular Biology* **4**. https://www.degruyter.com/document/doi/10.2202/1544-6115.1128/html (Accessed February 2, 2023).

Zhang Y, Castillo-Morales A, Jiang M, Zhu Y, Hu L, Urrutia AO, Kong X, Hurst LD. 2013. Genes That Escape X-Inactivation in Humans Have High Intraspecific Variability in Expression, Are Associated with Mental Impairment but Are Not Slow Evolving. *Mol Biol Evol* **30**: 2588–2601.

Zhang Z, Mathew D, Lim T, Huang S, Wherry EJ, Minn AJ, Ma Z, Zhang NR. 2023. Signal recovery in single cell batch integration. *bioRxiv* 2023.05.05.539614.

Zito A, Roberts AL, Visconti A, Rossi N, Andres-Ejarque R, Nardone S, Moustafa JES, Falchi M, Small KS. 2021. Escape from X-inactivation in twins exhibits intra- and inter-individual variability across tissues and is heritable. 2021.10.15.463586. https://www.biorxiv.org/content/10.1101/2021.10.15.463586v1 (Accessed April 21, 2022).

7. Appendix 1 – Supplementary tables

Supplemental Table S2.1. Escape annotations. Related to Figure 2.3.

Assigned labels of inactive, known escape, confident inactive, or novel escape for the 189 genes powered enough to investigate cross-tissue XCI escape.

gene	**escape_label**
OPHN1	confident inactive
ATRX	confident inactive
MAMLD1	confident inactive
RP11-1148L6.5	confident inactive
CXorf40B	confident inactive
WDR45	confident inactive
TREX2	confident inactive
ACOT9	confident inactive
ARMCX5-GPRASP2	confident inactive
EMD	confident inactive
PIR	confident inactive
PLXNA3	confident inactive
GPC4	confident inactive
DMD	confident inactive
RGN	confident inactive

HEPH	confident inactive
PRPS2	confident inactive
SRPK3	confident inactive
SLC6A8	confident inactive
TMEM47	confident inactive
EFNB1	confident inactive
MIR503HG	confident inactive
PNMA3	confident inactive
RP5-972B16.2	confident inactive
ZNF185	confident inactive
BEX2	confident inactive
SAT1	confident inactive
USP11	confident inactive
PDZD4	confident inactive
COL4A6	confident inactive
TCEAL2	confident inactive
TSPYL2	confident inactive
PLS3	confident inactive
BEX1	confident inactive
TSC22D3	confident inactive
PCSK1N	confident inactive
MSN	confident inactive
PLP1	confident inactive
SYN1	confident inactive
TMSB4X	confident inactive
ARSD	known escape
CA5B	known escape
CDK16	known escape
CXorf38	known escape
DDX3X	known escape
EIF1AX	known escape
EIF2S3	known escape
GEMIN8	known escape
GPM6B	known escape
GYG2	known escape
IQSEC2	known escape
KDM5C	known escape
KDM6A	known escape
MAOA	known escape

MED14	known escape
MSL3	known escape
MXRA5	known escape
NAP1L3	known escape
PLXNB3	known escape
PNPLA4	known escape
PRKX	known escape
RENBP	known escape
RP11-706O15.1	known escape
SMC1A	known escape
STS	known escape
SYAP1	known escape
SYTL4	known escape
TRAPPC2	known escape
TXLNG	known escape
UBA1	known escape
USP9X	known escape
XG	known escape
ZFX	known escape
ZRSR2	known escape
CTPS2	novel escape
CASK	novel escape
STARD8	novel escape
NLGN4X	novel escape
MECP2	novel escape
CLIC2	novel escape
F8	novel escape
PGK1	novel escape
SEPT6	novel escape
MPP1	novel escape
CXorf36	novel escape
ARHGAP4	novel escape
BTK	novel escape
COX7B	novel escape
RPL36A	novel escape
ITM2A	novel escape
CHRDL1	novel escape
VSIG4	novel escape
SASH3	novel escape

YIPF6	verified inactive
RLIM	verified inactive
TBC1D25	verified inactive
ZBTB33	verified inactive
PHF6	verified inactive
ZNF75D	verified inactive
THOC2	verified inactive
PHKA2	verified inactive
MORC4	verified inactive
GAB3	verified inactive
MTMR1	verified inactive
RPS6KA3	verified inactive
MED12	verified inactive
INTS6L	verified inactive
SCML1	verified inactive
WDR13	verified inactive
APEX2	verified inactive
ATP11C	verified inactive
ARHGAP6	verified inactive
ELF4	verified inactive
PHF8	verified inactive
TAF9B	verified inactive
MAP7D3	verified inactive
FAM122B	verified inactive
SLC25A43	verified inactive
SMIM10	verified inactive
ABCD1	verified inactive
RBBP7	verified inactive
DLG3	verified inactive
C1GALT1C1	verified inactive
SLC25A53	verified inactive
LINC01278	verified inactive
TBL1X	verified inactive
ZNF275	verified inactive
TMEM164	verified inactive
PRPS1	verified inactive
ARMCX1	verified inactive
FMR1	verified inactive
HUWE1	verified inactive

ZMAT1	verified inactive
AIFM1	verified inactive
SLC9A6	verified inactive
PJA1	verified inactive
HCFC1	verified inactive
NSDHL	verified inactive
FAM127C	verified inactive
OCRL	verified inactive
GRIPAP1	verified inactive
RBM10	verified inactive
APOO	verified inactive
ZDHHC9	verified inactive
ARMCX3	verified inactive
ZMYM3	verified inactive
UBL4A	verified inactive
GS1-358P8.4	verified inactive
EBP	verified inactive
REPS2	verified inactive
CLCN4	verified inactive
WWC3	verified inactive
LAGE3	verified inactive
DOCK11	verified inactive
DKC1	verified inactive
ARMCX2	verified inactive
SLC9A7	verified inactive
IL13RA1	verified inactive
VBP1	verified inactive
MAGEH1	verified inactive
TAZ	verified inactive
TSPAN6	verified inactive
HDAC6	verified inactive
SNX12	verified inactive
GABRE	verified inactive
TFE3	verified inactive
TSR2	verified inactive
HTATSF1	verified inactive
TIMM17B	verified inactive
LINC01420	verified inactive
FAM3A	verified inactive

RBMX	verified inactive
PDHA1	verified inactive
FAM127B	verified inactive
XIST	verified inactive
G6PD	verified inactive
PRAF2	verified inactive
CD99L2	verified inactive
BEX4	verified inactive
HNRNPH2	verified inactive
NONO	verified inactive
LDOC1	verified inactive
ATP6AP2	verified inactive
TCEAL3	verified inactive
MAGED1	verified inactive
MAOB	verified inactive
SH3BGRL	verified inactive
GDI1	verified inactive
IDS	verified inactive
ATP6AP1	verified inactive
MORF4L2	verified inactive
RBM3	verified inactive
MAGED2	verified inactive
TSIX	verified inactive
TCEAL4	verified inactive
SLC25A5	verified inactive

Supplemental Table S2.2. Germ layer-specific marker genes. Related to Figure 2.4.

Extracted germ layer-specific markers from the CIBERSORT deconvolution. Germ layer-specific markers are defined as genes that are identified through CIBERSORT as gene signatures exclusively for cell types of a single germ layer.

ensemblID	**name**	**germlayer_marker**	**tissue**	**celltype**
ENSG00000102125	TAZ	Mesoderm	breast	Immune..DC.macrophage. Endothelial.cell..lymphatic.
ENSG00000102287	GABRE	Mesoderm	breast	Adipocyte
ENSG00000133142	TCEAL4	Mesoderm	breast	Myoepithelial..basal.
ENSG00000203879	GDI1	Mesoderm	breast	Pericyte.SMC
ENSG00000013619	MAMLD1	Mesoderm	breast	Adipocyte
ENSG00000063601	MTMR1	Mesoderm	breast	Myoepithelial..basal.
ENSG00000125354	44810	Mesoderm	breast	Immune..DC.macrophage. Endothelial.cell..lymphatic.
ENSG00000131724	IL13RA1	Mesoderm	breast	Immune..DC.macrophage.
ENSG00000147065	MSN	Mesoderm	breast	Endothelial.cell..vascular. Pericyte.SMC
ENSG00000147113	CXorf36	Mesoderm	breast	Endothelial.cell..vascular. Endothelial.cell..lymphatic.
ENSG00000179222	MAGED1	Mesoderm	breast	Myoepithelial..basal.
ENSG00000185010	F8	Mesoderm	breast	Endothelial.cell..vascular.
ENSG00000204272	LINC01420	Mesoderm	breast	Myoepithelial..basal. Endothelial.cell..vascular.
ENSG00000123130	ACOT9	Mesoderm	breast	Immune..DC.macrophage.
ENSG00000155659	VSIG4	Mesoderm	breast	Immune..DC.macrophage.
ENSG00000157600	TMEM164	Mesoderm	breast	Adipocyte
ENSG00000102024	PLS3	Mesoderm	breast	Myoepithelial..basal. Endothelial.cell..vascular. Pericyte.SMC
ENSG00000158352	SHROOM4	Mesoderm	breast	Endothelial.cell..vascular.
ENSG00000102359	SRPX2	Mesoderm	breast	Adipocyte Fibroblast
ENSG00000147257	GPC3	Mesoderm	breast	Adipocyte Fibroblast
ENSG00000003096	KLHL13	Mesoderm	breast	Myoepithelial..basal.
ENSG00000158813	EDA	Mesoderm	breast	Immune..DC.macrophage.
ENSG00000047648	ARHGAP6	Mesoderm	breast	Immune..DC.macrophage. Fibroblast Pericyte.SMC
ENSG00000130150	MOSPD2	Mesoderm	esophagusMucosa	Immune..DC.
ENSG00000133142	TCEAL4	Mesoderm	esophagusMucosa	Myofibroblast
ENSG00000155659	VSIG4	Mesoderm	esophagusMucosa	Immune..DC.macrophage.
ENSG00000147113	CXorf36	Mesoderm	esophagusMucosa	Endothelial.cell..vascular. Endothelial.cell..lymphatic.

ENSG00000129675	ARHGEF6	Mesoderm	esophagusMucosa	Immune..DC.macrophage. Immune..T.cell. Immune..B.cell. Immune..DC. Immune..NK.cell. Immune..mast.cell.
ENSG00000165359	INTS6L	Mesoderm	esophagusMucosa	Immune..mast.cell.
ENSG00000197565	COL4A6	Mesoderm	esophagusMucosa	Myofibroblast
ENSG00000204272	LINC01420	Mesoderm	esophagusMucosa	Immune..DC. Immune..mast.cell.
ENSG00000123130	ACOT9	Mesoderm	esophagusMucosa	Myofibroblast Immune..DC.
ENSG00000181704	YIPF6	Mesoderm	esophagusMucosa	Immune..NK.cell.
ENSG00000155962	CLIC2	Mesoderm	esophagusMucosa	Immune..DC.
ENSG00000102119	EMD	Mesoderm	esophagusMucosa	Immune..NK.cell.
ENSG00000147251	DOCK11	Mesoderm	esophagusMucosa	Myofibroblast Endothelial.cell..vascular. Fibroblast Immune..DC.macrophage. Immune..T.cell. Immune..B.cell. Immune..DC. Immune..NK.cell. Immune..mast.cell.
ENSG00000125354	SEPT6	Mesoderm	esophagusMucosa	Myofibroblast Endothelial.cell..lymphatic. Immune..DC.macrophage. Immune..T.cell. Immune..B.cell. Immune..DC. Immune..NK.cell.
ENSG00000171100	MTM1	Mesoderm	esophagusMucosa	Myofibroblast Endothelial.cell..vascular. Immune..T.cell. Immune..mast.cell.
ENSG00000184205	TSPYL2	Mesoderm	esophagusMucosa	Myofibroblast
ENSG00000094631	HDAC6	Mesoderm	esophagusMucosa	Myofibroblast
ENSG00000069535	MAOB	Mesoderm	esophagusMucosa	Immune..mast.cell.
ENSG00000147010	SH3KBP1	Mesoderm	esophagusMucosa	Fibroblast Pericyte.SMC Immune..DC.macrophage. Immune..T.cell. Immune..B.cell. Immune..DC. Immune..NK.cell. Immune..mast.cell.
ENSG00000010671	BTK	Mesoderm	esophagusMucosa	Immune..B.cell. Immune..mast.cell.
ENSG00000158352	SHROOM4	Mesoderm	esophagusMucosa	Endothelial.cell..vascular.
ENSG00000047648	ARHGAP6	Mesoderm	esophagusMucosa	Myofibroblast Fibroblast Pericyte.SMC Immune..mast.cell.
ENSG00000122122	SASH3	Mesoderm	esophagusMucosa	Immune..NK.cell.
ENSG00000185010	F8	Mesoderm	esophagusMucosa	Endothelial.cell..lymphatic.
ENSG00000198814	GK	Mesoderm	esophagusMucosa	Immune..DC.
ENSG00000101974	ATP11C	Mesoderm	esophagusMucosa	Endothelial.cell..vascular. Immune..DC.macrophage. Immune..DC. Immune..NK.cell.
ENSG00000065923	SLC9A7	Mesoderm	esophagusMucosa	Immune..B.cell. Immune..DC.
ENSG00000160219	GAB3	Mesoderm	esophagusMucosa	Immune..DC.macrophage. Immune..T.cell. Immune..DC. Immune..NK.cell. Immune..mast.cell.
ENSG00000102221	JADE3	Mesoderm	esophagusMucosa	Immune..B.cell. Immune..DC.
ENSG00000101935	AMMECR1	Endoderm	esophagusMucosa	Epithelial.cell..suprabasal.
ENSG00000147140	NONO	Endoderm	esophagusMucosa	Epithelial.cell..basal.
ENSG00000147394	ZNF185	Endoderm	esophagusMucosa	Epithelial.cell..squamous. Epithelial.cell..suprabasal.
ENSG00000157625	TAB3	Endoderm	esophagusMucosa	Epithelial.cell..squamous.
ENSG00000184831	APOO	Endoderm	esophagusMucosa	Epithelial.cell..suprabasal. Epithelial.cell..basal.

ENSG00000082458	DLG3	Endoderm	esophagusMucosa	Mucous.cell
ENSG00000102287	GABRE	Endoderm	esophagusMucosa	Epithelial.cell..squamous. Epithelial.cell..suprabasal. Epithelial.cell..basal. Mucous.cell
ENSG00000130821	SLC6A8	Endoderm	esophagusMucosa	Epithelial.cell..squamous.
ENSG00000183337	BCOR	Endoderm	esophagusMucosa	Epithelial.cell..squamous.
ENSG00000102349	KLF8	Endoderm	esophagusMucosa	Epithelial.cell..squamous. Epithelial.cell..suprabasal.
ENSG00000071553	ATP6AP1	Endoderm	esophagusMucosa	Epithelial.cell..squamous.
ENSG00000182872	RBM10	Endoderm	esophagusMucosa	Epithelial.cell..squamous.
ENSG00000165591	FAAH2	Endoderm	esophagusMucosa	Epithelial.cell..suprabasal. Epithelial.cell..basal. Mucous.cell
ENSG00000172943	PHF8	Ectoderm	esophagusMucosa	Neuroendocrine
ENSG00000130827	PLXNA3	Ectoderm	esophagusMucosa	Neuroendocrine
ENSG00000071859	FAM50A	Ectoderm	esophagusMucosa	Neuroendocrine
ENSG00000146938	NLGN4X	Ectoderm	esophagusMucosa	Neuroendocrine Schwann.cell
ENSG00000269743	SLC25A53	Ectoderm	esophagusMucosa	Neuroendocrine
ENSG00000133131	MORC4	Ectoderm	esophagusMucosa	Neuroendocrine
ENSG00000067177	PHKA1	Ectoderm	esophagusMucosa	Neuroendocrine
ENSG00000158813	EDA	Ectoderm	esophagusMucosa	Schwann.cell
ENSG00000184343	SRPK3	Ectoderm	esophagusMucosa	Neuroendocrine
ENSG00000076716	GPC4	Mesoderm	esophagusMuscularis	Myocyte..smooth.muscle.
ENSG00000087842	PIR	Mesoderm	esophagusMuscularis	Endothelial.cell..vascular. ICCs
ENSG00000101849	TBL1X	Mesoderm	esophagusMuscularis	Myocyte..smooth.muscle. ICCs Pericyte.SMC
ENSG00000133142	TCEAL4	Mesoderm	esophagusMuscularis	Myocyte..smooth.muscle.
ENSG00000147113	CXorf36	Mesoderm	esophagusMuscularis	Endothelial.cell..vascular. Endothelial.cell..lymphatic.
ENSG00000155659	VSIG4	Mesoderm	esophagusMuscularis	Immune..DC.macrophage.
ENSG00000197565	COL4A6	Mesoderm	esophagusMuscularis	Myocyte..smooth.muscle.
ENSG00000131724	IL13RA1	Mesoderm	esophagusMuscularis	Immune..DC.macrophage.
ENSG00000147251	DOCK11	Mesoderm	esophagusMuscularis	Myocyte..smooth.muscle. Fibroblast Immune..DC.macrophage. Immune..mast.cell. Immune..T.cell.
ENSG00000165359	INTS6L	Mesoderm	esophagusMuscularis	ICCs
ENSG00000184205	TSPYL2	Mesoderm	esophagusMuscularis	Myocyte..smooth.muscle. ICCs
ENSG00000250349	RP5-972B16.2	Mesoderm	esophagusMuscularis	Endothelial.cell..vascular. Adipocyte Immune..B.cell.
ENSG00000086758	HUWE1	Mesoderm	esophagusMuscularis	ICCs Pericyte.SMC Adipocyte
ENSG00000188153	COL4A5	Mesoderm	esophagusMuscularis	Myocyte..smooth.muscle. ICCs Pericyte.SMC
ENSG00000204272	LINC01420	Mesoderm	esophagusMuscularis	Endothelial.cell..vascular. Immune..DC.macrophage.
ENSG00000185010	F8	Mesoderm	esophagusMuscularis	Endothelial.cell..vascular. Endothelial.cell..lymphatic. Adipocyte

ENSG00000131171	SH3BGRL	Mesoderm	esophagusMuscularis	Myocyte..smooth.muscle. Immune..DC.macrophage. Pericyte.SMC
ENSG00000129675	ARHGEF6	Mesoderm	esophagusMuscularis	Immune..mast.cell.
ENSG00000157600	TMEM164	Mesoderm	esophagusMuscularis	Immune..mast.cell.
ENSG00000122122	SASH3	Mesoderm	esophagusMuscularis	Immune..B.cell.
ENSG00000158352	SHROOM4	Mesoderm	esophagusMuscularis	Endothelial.cell..vascular.
ENSG00000173698	ADGRG2	Mesoderm	esophagusMuscularis	ICCs Adipocyte
ENSG00000165675	ENOX2	Mesoderm	esophagusMuscularis	Fibroblast Immune..DC.macrophage. Adipocyte
ENSG00000171100	MTM1	Mesoderm	esophagusMuscularis	Immune..DC.macrophage.
ENSG00000101871	MID1	Mesoderm	esophagusMuscularis	Myocyte..smooth.muscle. Endothelial.cell..lymphatic. ICCs Fibroblast
ENSG00000010404	IDS	Ectoderm	esophagusMuscularis	Neuronal
ENSG00000123560	PLP1	Ectoderm	esophagusMuscularis	Neuronal Schwann.cell
ENSG00000126970	ZC4H2	Ectoderm	esophagusMuscularis	Neuronal
ENSG00000102109	PCSK1N	Ectoderm	esophagusMuscularis	Neuronal
ENSG00000198689	SLC9A6	Ectoderm	esophagusMuscularis	Neuronal
ENSG00000146938	NLGN4X	Ectoderm	esophagusMuscularis	Neuronal Schwann.cell
ENSG00000067177	PHKA1	Ectoderm	esophagusMuscularis	Neuronal
ENSG00000133169	BEX1	Ectoderm	esophagusMuscularis	Neuronal
ENSG00000087842	PIR	Mesoderm	heartAtrialAppendage	Myocyte..cardiac. Endothelial.cell..vascular. Fibroblast Adipocyte
ENSG00000102144	PGK1	Mesoderm	heartAtrialAppendage	Myocyte..cardiac..cytoplasmic.
ENSG00000102287	GABRE	Mesoderm	heartAtrialAppendage	Adipocyte
ENSG00000102349	KLF8	Mesoderm	heartAtrialAppendage	Adipocyte
ENSG00000125354	44810	Mesoderm	heartAtrialAppendage	Immune..DC.macrophage. Immune..B.cell. Immune..T.cell. Immune..NK.cell.
ENSG00000130821	SLC6A8	Mesoderm	heartAtrialAppendage	Myocyte..cardiac..cytoplasmic. Myocyte..cardiac.
ENSG00000131171	SH3BGRL	Mesoderm	heartAtrialAppendage	Immune..mast.cell. Pericyte.SMC Immune..DC.macrophage. Adipocyte Immune..B.cell.
ENSG00000131174	COX7B	Mesoderm	heartAtrialAppendage	Myocyte..cardiac..cytoplasmic.
ENSG00000131724	IL13RA1	Mesoderm	heartAtrialAppendage	Immune..DC.macrophage.
ENSG00000147119	CHST7	Mesoderm	heartAtrialAppendage	Fibroblast
ENSG00000165359	INTS6L	Mesoderm	heartAtrialAppendage	Immune..mast.cell.
ENSG00000185010	F8	Mesoderm	heartAtrialAppendage	Endothelial.cell..vascular. Endothelial.cell..lymphatic.
ENSG00000204272	LINC01420	Mesoderm	heartAtrialAppendage	Myocyte..cardiac. Endothelial.cell..vascular. Adipocyte Immune..B.cell.
ENSG00000078596	ITM2A	Mesoderm	heartAtrialAppendage	Endothelial.cell..vascular.
ENSG00000102125	TAZ	Mesoderm	heartAtrialAppendage	Myocyte..cardiac.
ENSG00000147113	CXorf36	Mesoderm	heartAtrialAppendage	Endothelial.cell..vascular. Endothelial.cell..lymphatic.

ENSG00000044446	PHKA2	Mesoderm	heartAtrialAppendage	Myocyte..cardiac.
ENSG00000101849	TBL1X	Mesoderm	heartAtrialAppendage	Myocyte..cardiac. Pericyte.SMC Immune..DC.macrophage. Adipocyte Immune..B.cell. Immune..T.cell. Immune..NK.cell.
ENSG00000157600	TMEM164	Mesoderm	heartAtrialAppendage	Myocyte..cardiac. Immune..mast.cell. Immune..DC.macrophage. Adipocyte
ENSG00000160219	GAB3	Mesoderm	heartAtrialAppendage	Myocyte..cardiac. Immune..mast.cell. Immune..B.cell.
ENSG00000184831	APOO	Mesoderm	heartAtrialAppendage	Myocyte..cardiac..cytoplasmic. Adipocyte
ENSG00000102317	RBM3	Mesoderm	heartAtrialAppendage	Myocyte..cardiac. Adipocyte
ENSG00000131828	PDHA1	Mesoderm	heartAtrialAppendage	Myocyte..cardiac..cytoplasmic.
ENSG00000068366	ACSL4	Mesoderm	heartAtrialAppendage	Myocyte..cardiac. Immune..mast.cell. Immune..DC.macrophage. Adipocyte Immune..B.cell.
ENSG00000102024	PLS3	Mesoderm	heartAtrialAppendage	Myocyte..cardiac. Endothelial.cell..vascular. Fibroblast Pericyte.SMC Endothelial.cell..lymphatic. Adipocyte
ENSG00000147251	DOCK11	Mesoderm	heartAtrialAppendage	Immune..mast.cell. Immune..DC.macrophage. Adipocyte Immune..B.cell. Immune..T.cell. Immune..NK.cell.
ENSG00000147257	GPC3	Mesoderm	heartAtrialAppendage	Adipocyte
ENSG00000169083	AR	Mesoderm	heartAtrialAppendage	Myocyte..cardiac. Adipocyte
ENSG00000185825	BCAP31	Mesoderm	heartAtrialAppendage	Myocyte..cardiac..cytoplasmic. Endothelial.cell..lymphatic. Adipocyte
ENSG00000155659	VSIG4	Mesoderm	heartAtrialAppendage	Immune..DC.macrophage.
ENSG00000165192	ASB11	Mesoderm	heartAtrialAppendage	Myocyte..cardiac.
ENSG00000169057	MECP2	Mesoderm	heartAtrialAppendage	Myocyte..cardiac. Immune..mast.cell. Endothelial.cell..lymphatic. Adipocyte Immune..T.cell. Immune..NK.cell.
ENSG00000125351	UPF3B	Mesoderm	heartAtrialAppendage	Myocyte..cardiac. Immune..mast.cell.
ENSG00000147123	NDUFB11	Mesoderm	heartAtrialAppendage	Myocyte..cardiac..cytoplasmic.
ENSG00000122122	SASH3	Mesoderm	heartAtrialAppendage	Immune..B.cell. Immune..NK.cell.
ENSG00000196998	WDR45	Mesoderm	heartAtrialAppendage	Immune..B.cell.
ENSG00000125356	NDUFA1	Mesoderm	heartAtrialAppendage	Myocyte..cardiac..cytoplasmic.
ENSG00000091482	SMPX	Mesoderm	heartAtrialAppendage	Myocyte..cardiac..cytoplasmic.
ENSG00000131725	WDR44	Mesoderm	heartAtrialAppendage	Myocyte..cardiac. Immune..B.cell. Immune..T.cell.
ENSG00000047634	SCML1	Mesoderm	heartAtrialAppendage	Myocyte..cardiac. Immune..mast.cell. Immune..DC.macrophage. Adipocyte
ENSG00000134590	FAM127A	Mesoderm	heartAtrialAppendage	Myocyte..cardiac..cytoplasmic. Adipocyte
ENSG00000166681	BEX3	Mesoderm	heartAtrialAppendage	Myocyte..cardiac..cytoplasmic.
ENSG00000188419	CHM	Mesoderm	heartAtrialAppendage	Myocyte..cardiac. Immune..mast.cell.
ENSG00000086758	HUWE1	Mesoderm	heartAtrialAppendage	Myocyte..cardiac..cytoplasmic. Myocyte..cardiac. Immune..mast.cell. Endothelial.cell..lymphatic. Adipocyte Immune..T.cell.
ENSG00000101871	MID1	Mesoderm	heartAtrialAppendage	Myocyte..cardiac. Fibroblast Adipocyte
ENSG00000198814	GK	Mesoderm	heartAtrialAppendage	Myocyte..cardiac.

ENSG00000008086	CDKL5	Mesoderm	heartAtrialAppendage	Myocyte..cardiac. Immune..B.cell.
ENSG00000158352	SHROOM4	Mesoderm	heartAtrialAppendage	Endothelial.cell..vascular.
ENSG00000169891	REPS2	Mesoderm	heartAtrialAppendage	Endothelial.cell..lymphatic.
ENSG00000047648	ARHGAP6	Mesoderm	heartAtrialAppendage	Immune..mast.cell. Fibroblast Pericyte.SMC Endothelial.cell..lymphatic. Immune..DC.macrophage.
ENSG00000123572	NRK	Mesoderm	heartAtrialAppendage	Fibroblast
ENSG00000188153	COL4A5	Mesoderm	heartAtrialAppendage	Myocyte..cardiac.
ENSG00000156531	PHF6	Mesoderm	heartAtrialAppendage	Immune..B.cell.
ENSG00000197779	ZNF81	Mesoderm	heartAtrialAppendage	Myocyte..cardiac.
ENSG00000123560	PLP1	Ectoderm	heartAtrialAppendage	Schwann.cell
ENSG00000078596	ITM2A	Mesoderm	heartLeftVentricle	Endothelial.cell..vascular.
ENSG00000087842	PIR	Mesoderm	heartLeftVentricle	Myocyte..cardiac. Endothelial.cell..vascular. Fibroblast Adipocyte
ENSG00000091482	SMPX	Mesoderm	heartLeftVentricle	Myocyte..cardiac..cytoplasmic.
ENSG00000102125	TAZ	Mesoderm	heartLeftVentricle	Myocyte..cardiac.
ENSG00000102287	GABRE	Mesoderm	heartLeftVentricle	Adipocyte
ENSG00000130821	SLC6A8	Mesoderm	heartLeftVentricle	Myocyte..cardiac..cytoplasmic. Myocyte..cardiac.
ENSG00000131725	WDR44	Mesoderm	heartLeftVentricle	Myocyte..cardiac. Immune..B.cell. Immune..T.cell.
ENSG00000147113	CXorf36	Mesoderm	heartLeftVentricle	Endothelial.cell..vascular. Endothelial.cell..lymphatic.
ENSG00000157600	TMEM164	Mesoderm	heartLeftVentricle	Myocyte..cardiac. Immune..mast.cell. Immune..DC.macrophage. Adipocyte
ENSG00000160219	GAB3	Mesoderm	heartLeftVentricle	Myocyte..cardiac. Immune..mast.cell. Immune..B.cell.
ENSG00000184831	APOO	Mesoderm	heartLeftVentricle	Myocyte..cardiac..cytoplasmic. Adipocyte
ENSG00000185010	F8	Mesoderm	heartLeftVentricle	Endothelial.cell..vascular. Endothelial.cell..lymphatic.
ENSG00000102317	RBM3	Mesoderm	heartLeftVentricle	Myocyte..cardiac. Adipocyte
ENSG00000131828	PDHA1	Mesoderm	heartLeftVentricle	Myocyte..cardiac..cytoplasmic.
ENSG00000155659	VSIG4	Mesoderm	heartLeftVentricle	Immune..DC.macrophage.
ENSG00000086758	HUWE1	Mesoderm	heartLeftVentricle	Myocyte..cardiac..cytoplasmic. Myocyte..cardiac. Immune..mast.cell. Endothelial.cell..lymphatic. Adipocyte Immune..T.cell.
ENSG00000147123	NDUFB11	Mesoderm	heartLeftVentricle	Myocyte..cardiac..cytoplasmic.
ENSG00000165192	ASB11	Mesoderm	heartLeftVentricle	Myocyte..cardiac.
ENSG00000185825	BCAP31	Mesoderm	heartLeftVentricle	Myocyte..cardiac..cytoplasmic. Endothelial.cell..lymphatic. Adipocyte
ENSG00000204272	LINC01420	Mesoderm	heartLeftVentricle	Myocyte..cardiac. Endothelial.cell..vascular. Adipocyte Immune..B.cell.
ENSG00000125356	NDUFA1	Mesoderm	heartLeftVentricle	Myocyte..cardiac..cytoplasmic.
ENSG00000131171	SH3BGRL	Mesoderm	heartLeftVentricle	Immune..mast.cell. Pericyte.SMC Immune..DC.macrophage. Adipocyte Immune..B.cell.
ENSG00000131724	IL13RA1	Mesoderm	heartLeftVentricle	Immune..DC.macrophage.

ENSG00000101849	TBL1X	Mesoderm	heartLeftVentricle	Myocyte..cardiac. Pericyte.SMC Immune..DC.macrophage. Adipocyte Immune..B.cell. Immune..T.cell. Immune..NK.cell.
ENSG00000102144	PGK1	Mesoderm	heartLeftVentricle	Myocyte..cardiac..cytoplasmic.
ENSG00000131174	COX7B	Mesoderm	heartLeftVentricle	Myocyte..cardiac..cytoplasmic.
ENSG00000169083	AR	Mesoderm	heartLeftVentricle	Myocyte..cardiac. Adipocyte
ENSG00000147251	DOCK11	Mesoderm	heartLeftVentricle	Immune..mast.cell. Immune..DC.macrophage. Adipocyte Immune..B.cell. Immune..T.cell. Immune..NK.cell.
ENSG00000169057	MECP2	Mesoderm	heartLeftVentricle	Myocyte..cardiac. Immune..mast.cell. Endothelial.cell..lymphatic. Adipocyte Immune..T.cell. Immune..NK.cell.
ENSG00000196998	WDR45	Mesoderm	heartLeftVentricle	Immune..B.cell.
ENSG00000102024	PLS3	Mesoderm	heartLeftVentricle	Myocyte..cardiac. Endothelial.cell..vascular. Fibroblast Pericyte.SMC Endothelial.cell..lymphatic. Adipocyte
ENSG00000158352	SHROOM4	Mesoderm	heartLeftVentricle	Endothelial.cell..vascular.
ENSG00000044446	PHKA2	Mesoderm	heartLeftVentricle	Myocyte..cardiac.
ENSG00000125354	44810	Mesoderm	heartLeftVentricle	Immune..DC.macrophage. Immune..B.cell. Immune..T.cell. Immune..NK.cell.
ENSG00000068366	ACSL4	Mesoderm	heartLeftVentricle	Myocyte..cardiac. Immune..mast.cell. Immune..DC.macrophage. Adipocyte Immune..B.cell.
ENSG00000147119	CHST7	Mesoderm	heartLeftVentricle	Fibroblast
ENSG00000166681	BEX3	Mesoderm	heartLeftVentricle	Myocyte..cardiac..cytoplasmic.
ENSG00000188419	CHM	Mesoderm	heartLeftVentricle	Myocyte..cardiac. Immune..mast.cell.
ENSG00000125351	UPF3B	Mesoderm	heartLeftVentricle	Myocyte..cardiac. Immune..mast.cell.
ENSG00000122122	SASH3	Mesoderm	heartLeftVentricle	Immune..B.cell. Immune..NK.cell.
ENSG00000134590	FAM127A	Mesoderm	heartLeftVentricle	Myocyte..cardiac..cytoplasmic. Adipocyte
ENSG00000047634	SCML1	Mesoderm	heartLeftVentricle	Myocyte..cardiac. Immune..mast.cell. Immune..DC.macrophage. Adipocyte
ENSG00000156531	PHF6	Mesoderm	heartLeftVentricle	Immune..B.cell.
ENSG00000165359	INTS6L	Mesoderm	heartLeftVentricle	Immune..mast.cell.
ENSG00000101871	MID1	Mesoderm	heartLeftVentricle	Myocyte..cardiac. Fibroblast Adipocyte
ENSG00000123572	NRK	Mesoderm	heartLeftVentricle	Fibroblast
ENSG00000197779	ZNF81	Mesoderm	heartLeftVentricle	Myocyte..cardiac.
ENSG00000198814	GK	Mesoderm	heartLeftVentricle	Myocyte..cardiac.
ENSG00000165197	VEGFD	Mesoderm	heartLeftVentricle	Myocyte..cardiac. Fibroblast
ENSG00000147257	GPC3	Mesoderm	heartLeftVentricle	Adipocyte
ENSG00000171100	MTM1	Mesoderm	heartLeftVentricle	Immune..DC.macrophage. Immune..B.cell. Immune..T.cell.
ENSG00000123560	PLP1	Ectoderm	heartLeftVentricle	Schwann.cell
ENSG00000068400	GRIPAP1	Mesoderm	lung	Immune..NK.cell.
ENSG00000069535	MAOB	Mesoderm	lung	Fibroblast Immune..mast.cell.

ENSG00000089820	ARHGAP4	Mesoderm	lung	Immune..mast.cell. Immune..B.cell.
ENSG00000102144	PGK1	Mesoderm	lung	Immune..DC.macrophage.
ENSG00000122122	SASH3	Mesoderm	lung	Immune..mast.cell. Immune..B.cell.
ENSG00000147113	CXorf36	Mesoderm	lung	Endothelial.cell..vascular. Endothelial.cell..lymphatic.
ENSG00000160219	GAB3	Mesoderm	lung	Immune..alveolar.macrophage. Immune..mast.cell.
ENSG00000250349	RP5-972B16.2	Mesoderm	lung	Endothelial.cell..vascular. Endothelial.cell..lymphatic.
ENSG00000047648	ARHGAP6	Mesoderm	lung	Immune..alveolar.macrophage. Pericyte.SMC Fibroblast Immune..mast.cell. Immune..DC.macrophage.
ENSG00000101974	ATP11C	Mesoderm	lung	Endothelial.cell..vascular. Endothelial.cell..lymphatic. Immune..NK.cell.
ENSG00000102359	SRPX2	Mesoderm	lung	Fibroblast
ENSG00000125354	44810	Mesoderm	lung	Immune..alveolar.macrophage. Pericyte.SMC Immune..mast.cell. Immune..T.cell. Immune..DC.macrophage. Immune..B.cell. Immune..NK.cell.
ENSG00000198814	GK	Mesoderm	lung	Immune..alveolar.macrophage. Immune..mast.cell. Immune..DC.macrophage.
ENSG00000204272	LINC01420	Mesoderm	lung	Immune..alveolar.macrophage. Immune..mast.cell.
ENSG00000133142	TCEAL4	Mesoderm	lung	Pericyte.SMC Fibroblast
ENSG00000165197	VEGFD	Mesoderm	lung	Fibroblast
ENSG00000131724	IL13RA1	Mesoderm	lung	Immune..alveolar.macrophage. Immune..DC.macrophage.
ENSG00000147168	IL2RG	Mesoderm	lung	Immune..NK.cell.
ENSG00000185010	F8	Mesoderm	lung	Endothelial.cell..vascular. Endothelial.cell..lymphatic.
ENSG00000155659	VSIG4	Mesoderm	lung	Immune..alveolar.macrophage. Immune..DC.macrophage.
ENSG00000102172	SMS	Mesoderm	lung	Immune..DC.macrophage.
ENSG00000130988	RGN	Mesoderm	lung	Pericyte.SMC
ENSG00000184194	GPR173	Mesoderm	lung	Fibroblast
ENSG00000102096	PIM2	Mesoderm	lung	Immune..B.cell.
ENSG00000155966	AFF2	Mesoderm	lung	Pericyte.SMC Immune..mast.cell. Immune..B.cell.
ENSG00000158813	EDA	Mesoderm	lung	Fibroblast Immune..DC.macrophage.
ENSG00000165675	ENOX2	Mesoderm	lung	Immune..alveolar.macrophage. Immune..DC.macrophage. Immune..B.cell.
ENSG00000129675	ARHGEF6	Mesoderm	lung	Immune..alveolar.macrophage. Immune..mast.cell. Immune..T.cell.
ENSG00000076716	GPC4	Endoderm	lung	Epithelial.cell..alveolar.type.II. Epithelial.cell..club. Epithelial.cell..alveolar.type.I. Epithelial.cell..basal.
ENSG00000101940	WDR13	Endoderm	lung	Epithelial.cell..ciliated.
ENSG00000102181	CD99L2	Endoderm	lung	Epithelial.cell..basal.
ENSG00000008086	CDKL5	Endoderm	lung	Epithelial.cell..alveolar.type.II. Epithelial.cell..club. Epithelial.cell..alveolar.type.I. Epithelial.cell..basal.

ENSG00000181704	YIPF6	Endoderm	lung	Epithelial.cell..basal.
ENSG00000102316	MAGED2	Endoderm	lung	Epithelial.cell..basal.
ENSG00000029993	HMGB3	Endoderm	lung	Epithelial.cell..basal.
ENSG00000183337	BCOR	Endoderm	lung	Epithelial.cell..alveolar.type.II.
ENSG00000197565	COL4A6	Endoderm	lung	Epithelial.cell..basal.
ENSG00000165164	CFAP47	Endoderm	lung	Epithelial.cell..ciliated.
ENSG00000089682	RBM41	Endoderm	lung	Epithelial.cell..ciliated.
ENSG00000129682	FGF13	Endoderm	lung	Epithelial.cell..club.
ENSG00000069535	MAOB	Mesoderm	skeletalMuscle	Adipocyte
ENSG00000087842	PIR	Mesoderm	skeletalMuscle	Endothelial.cell..vascular. Myocyte..sk Adipocyte Myocyte..NMJ.rich.
ENSG00000147113	CXorf36	Mesoderm	skeletalMuscle	Endothelial.cell..vascular. Endothelial.cell..lymphatic.
ENSG00000157600	TMEM164	Mesoderm	skeletalMuscle	Myocyte..sk Adipocyte Myocyte..NMJ.rich.
ENSG00000131724	IL13RA1	Mesoderm	skeletalMuscle	Immune..DC.macrophage.
ENSG00000155659	VSIG4	Mesoderm	skeletalMuscle	Immune..DC.macrophage.
ENSG00000250349	RP5-972B16.2	Mesoderm	skeletalMuscle	Endothelial.cell..lymphatic.
ENSG00000078596	ITM2A	Mesoderm	skeletalMuscle	Immune..T.cell.
ENSG00000102024	PLS3	Mesoderm	skeletalMuscle	Endothelial.cell..vascular. Pericyte.SMC Endothelial.cell..lymphatic.
ENSG00000185010	F8	Mesoderm	skeletalMuscle	Endothelial.cell..vascular. Endothelial.cell..lymphatic.
ENSG00000175556	LONRF3	Mesoderm	skeletalMuscle	Myocyte..NMJ.rich.
ENSG00000067177	PHKA1	Mesoderm	skeletalMuscle	Myocyte..sk Myocyte..NMJ.rich.
ENSG00000129675	ARHGEF6	Mesoderm	skeletalMuscle	Immune..T.cell. Myocyte..NMJ.rich.
ENSG00000165197	VEGFD	Mesoderm	skeletalMuscle	Myocyte..sk Myocyte..NMJ.rich.
ENSG00000179222	MAGED1	Mesoderm	skeletalMuscle	Adipocyte Immune..T.cell. Immune..DC.macrophage.
ENSG00000011201	ANOS1	Mesoderm	skeletalMuscle	Fibroblast
ENSG00000165175	MID1IP1	Mesoderm	skeletalMuscle	Fibroblast
ENSG00000169083	AR	Mesoderm	skeletalMuscle	Fibroblast Adipocyte
ENSG00000188153	COL4A5	Mesoderm	skeletalMuscle	Myocyte..sk Pericyte.SMC Myocyte..NMJ.rich.
ENSG00000125351	UPF3B	Mesoderm	skeletalMuscle	Endothelial.cell..vascular.
ENSG00000131171	SH3BGRL	Mesoderm	skeletalMuscle	Pericyte.SMC Immune..DC.macrophage.
ENSG00000102287	GABRE	Mesoderm	skeletalMuscle	Adipocyte
ENSG00000197565	COL4A6	Mesoderm	skeletalMuscle	Myocyte..NMJ.rich.
ENSG00000179542	SLITRK4	Mesoderm	skeletalMuscle	Fibroblast Myocyte..NMJ.rich.
ENSG00000129682	FGF13	Mesoderm	skeletalMuscle	Myocyte..sk Adipocyte Satellite.cell Myocyte..NMJ.rich.
ENSG00000147010	SH3KBP1	Mesoderm	skeletalMuscle	Adipocyte Immune..T.cell. Immune..DC.macrophage. Immune..NK.cell.

ENSG00000068366	ACSL4	Mesoderm	skeletalMuscle	Immune..DC.macrophage.
ENSG00000171365	CLCN5	Mesoderm	skeletalMuscle	Myocyte..sk Satellite.cell Myocyte..NMJ.rich.
ENSG00000078061	ARAF	Mesoderm	skeletalMuscle	Immune..T.cell.
ENSG00000102359	SRPX2	Mesoderm	skeletalMuscle	Fibroblast Adipocyte Satellite.cell
ENSG00000185222	TCEAL9	Mesoderm	skeletalMuscle	Fibroblast
ENSG00000101868	POLA1	Mesoderm	skeletalMuscle	Immune..T.cell. Myocyte..NMJ.rich.
ENSG00000158813	EDA	Mesoderm	skeletalMuscle	Satellite.cell Immune..DC.macrophage.
ENSG00000083750	RRAGB	Mesoderm	skeletalMuscle	Fibroblast Satellite.cell
ENSG00000077713	SLC25A43	Mesoderm	skeletalMuscle	Adipocyte
ENSG00000101901	ALG13	Mesoderm	skeletalMuscle	Endothelial.cell..lymphatic.
ENSG00000147251	DOCK11	Mesoderm	skeletalMuscle	Fibroblast Pericyte.SMC Adipocyte Immune..DC.macrophage. Immune..NK.cell.
ENSG00000203950	FAM127B	Ectoderm	skeletalMuscle	Schwann.cell
ENSG00000013619	MAMLD1	Mesoderm	skinLowerLeg	Adipocyte
ENSG00000102024	PLS3	Mesoderm	skinLowerLeg	Unknown Endothelial.cell..vascular.
ENSG00000147065	MSN	Mesoderm	skinLowerLeg	Unknown Adipocyte Endothelial.cell..vascular.
ENSG00000157600	TMEM164	Mesoderm	skinLowerLeg	Adipocyte
ENSG00000171365	CLCN5	Mesoderm	skinLowerLeg	Endothelial.cell..vascular.
ENSG00000185222	TCEAL9	Mesoderm	skinLowerLeg	Unknown Pericyte.SMC Adipocyte Endothelial.cell..vascular.
ENSG00000205542	TMSB4X	Mesoderm	skinLowerLeg	Unknown Endothelial.cell..vascular.
ENSG00000102287	GABRE	Mesoderm	skinLowerLeg	Adipocyte Fibroblast
ENSG00000155962	CLIC2	Mesoderm	skinLowerLeg	Endothelial.cell..vascular.
ENSG00000182872	RBM10	Mesoderm	skinLowerLeg	Unknown Pericyte.SMC Adipocyte Fibroblast
ENSG00000147251	DOCK11	Mesoderm	skinLowerLeg	Adipocyte Fibroblast
ENSG00000169057	MECP2	Mesoderm	skinLowerLeg	Pericyte.SMC Adipocyte Endothelial.cell..vascular.
ENSG00000158352	SHROOM4	Mesoderm	skinLowerLeg	Endothelial.cell..vascular.
ENSG00000165591	FAAH2	Ectoderm	skinLowerLeg	Sweat.gland.cell Epithelial.cell..basal.keratinocyte.
ENSG00000047230	CTPS2	Ectoderm	skinLowerLeg	Melanocyte
ENSG00000131171	SH3BGRL	Ectoderm	skinLowerLeg	Melanocyte
ENSG00000123560	PLP1	Ectoderm	skinLowerLeg	Melanocyte
ENSG00000101928	MOSPD1	Ectoderm	skinLowerLeg	Melanocyte
ENSG00000101850	GPR143	Ectoderm	skinLowerLeg	Melanocyte
ENSG00000102287	GABRE	Mesoderm	skinSuprapubic	Adipocyte Fibroblast
ENSG00000157600	TMEM164	Mesoderm	skinSuprapubic	Adipocyte
ENSG00000205542	TMSB4X	Mesoderm	skinSuprapubic	Unknown Endothelial.cell..vascular.

ENSG00000147251	DOCK11	Mesoderm	skinSuprapubic	Adipocyte Fibroblast
ENSG00000182872	RBM10	Mesoderm	skinSuprapubic	Unknown Pericyte.SMC Adipocyte Fibroblast
ENSG00000147065	MSN	Mesoderm	skinSuprapubic	Unknown Adipocyte Endothelial.cell..vascular.
ENSG00000169057	MECP2	Mesoderm	skinSuprapubic	Pericyte.SMC Adipocyte Endothelial.cell..vascular.
ENSG00000013619	MAMLD1	Mesoderm	skinSuprapubic	Adipocyte
ENSG00000102024	PLS3	Mesoderm	skinSuprapubic	Unknown Endothelial.cell..vascular.
ENSG00000185222	TCEAL9	Mesoderm	skinSuprapubic	Unknown Pericyte.SMC Adipocyte Endothelial.cell..vascular.
ENSG00000155962	CLIC2	Mesoderm	skinSuprapubic	Endothelial.cell..vascular.
ENSG00000171365	CLCN5	Mesoderm	skinSuprapubic	Endothelial.cell..vascular.
ENSG00000047230	CTPS2	Ectoderm	skinSuprapubic	Melanocyte
ENSG00000131171	SH3BGRL	Ectoderm	skinSuprapubic	Melanocyte
ENSG00000165591	FAAH2	Ectoderm	skinSuprapubic	Sweat.gland.cell Epithelial.cell..basal.keratinocyte.
ENSG00000123560	PLP1	Ectoderm	skinSuprapubic	Melanocyte
ENSG00000101928	MOSPD1	Ectoderm	skinSuprapubic	Melanocyte
ENSG00000101850	GPR143	Ectoderm	skinSuprapubic	Melanocyte

Supplemental Table S4.1

Primary tissue and neural organoid dataset download and metadata

Table containing the study origin and download links for all primary tissue and organoid scRNA-seq datasets. The batch variable column details the meta-data used in determining batch. The region/protocol column details the sampled primary tissue brain regions or the organoid differentiation protocol.

Column1	**Study**	**Batch variable**	**Dataset**	**Region/protocol**	**Download**
annotated fetal	Polioudakis et. al.	Library	Geschwind (GW17-18)	cortical anlage	http://solo.bmap.ucla.edu/shiny/webapp/
			Plath (GW17-18)	cortical anlage	
	Fan et al.	NA	GW7-28	cerebral cortex and pons	GEO: GSE120046, GSE120046_brain_all_UMIcounts.txt, GSE120046_metadata.txt
	Bhaduri et al.	Donor/Age	GW14	Multi-region	NeMO Archive RRID: SCR_002001, Metadata from Supp. Table 1 of publication
			GW18_2	Multi-region	
			GW19	Multi-region	
			GW19_2	Multi-region	
			GW20	Multi-region	
			GW20_31	Multi-region	
			GW20_34	Multi-region	
			GW25	Multi-region	
	Braun et al.	Donor/Age	XDD:348 (GW5)	Multi-region	https://github.com/linnarsson-lab/developing-human-brain/
			XDD:400 (GW5.5)	Multi-region	
			XDD:326 (GW6)	Multi-region	
			XDD:395 (GW6)	Multi-region	
			BRC2073 (GW6.6)	Multi-region	
			BRC2106A (GW6.6)	Multi-region	
			BRC2147 (GW6.7)	Multi-region	
			BRC2061 (GW6.9)	Multi-region	
			BRC2110 (GW6.9)	Multi-region	
			BRC2114 (GW6.9)	Multi-region	
			BRC2191 (GW6.9)	Multi-region	
			XDD:398 (GW7)	Multi-region	
			XHU:305 (GW7.5)	Multi-region	
			BRC2006 (GW8)	Multi-region	
			BRC2021 (GW8)	Multi-region	

			XDD:334 (GW8)	Multi-region	
			BRC2057 (GW8.1)	Multi-region	
			XDD:313 (GW8.5)	Multi-region	
			XDD:342 (GW8.5)	Multi-region	
			XHU:307 (GW9.2)	Multi-region	
			XHU:292 (GW9.5)	Multi-region	
			XHU:297 (GW10)	Multi-region	
			XDD:358 (GW11.5)	Multi-region	
			XDD:351 (GW12)	Multi-region	
			XDD:359 (GW13	Multi-region	
			XDD:385 (GW14)	Multi-region	
unannotated fetal	Shi et al.	NA	NA	ganglionic eminences/subpallium	GEO: GSE135827, GSE135827_GE_mat_raw_count_with_week_info.txt
	Zhou et al.	NA	NA	hypothalamus	GEO: GSE169109
	Yu et al.	Donor/Age	GSM5032680 (GW9)	ganglionic eminences/subpallium	GEO: GSE165388
			GSM5032681 (GW10)	ganglionic eminences/subpallium	
			GSM5032682 (GW11)	ganglionic eminences/subpallium	
			GSM5032683 (GW12)	ganglionic eminences/subpallium	
	Trevino et al.	Donor/Age	GW16	cerebral cortex	GEO: GSE162170, GSE162170_rna_counts.tsv.gz, GSE162170_rna_cell_metadata.txt.gz
			GW20	cerebral cortex	
			GW21	cerebral cortex	
			GW24	cerebral cortex	
	Bhaduri et al.	Donor/Age	GW16	Multi-region	NeMO Archive RRID: SCR_002001, Metadata from Supp. Table 1 of publication
			GW18	Multi-region	
			GW20	Multi-region	
			GW22T	Multi-region	
adult	Jorstad et al.	NA	NA	Medial temporal gyrus	from collaborators
organoid	Uzquiano et al.	Dataset/Cell line	Mito210c1 (23 days)	dorsal patterned forebrain	https://singlecell.broadinstitute.org/single_cell/study/SCP1756/cortical-organoids-atlas
			PGP1 (23 days)	dorsal patterned forebrain	
			dataset 1 (1 month)	dorsal patterned forebrain	
			dataset 2 (1 month)	dorsal patterned forebrain	
			dataset 3 (1 month)	dorsal patterned forebrain	
			dataset 4 (1 month)	dorsal patterned forebrain	

			Mito210c1 (1.5 months)	dorsal patterned forebrain	
			PGP1 (1.5 months)	dorsal patterned forebrain	
			Mito210c1 (2 months)	dorsal patterned forebrain	
			PGP1 (2 months)	dorsal patterned forebrain	
			dataset 1 (3 months)	dorsal patterned forebrain	
			dataset 2 (3 months)	dorsal patterned forebrain	
			dataset 3 (3 months)	dorsal patterned forebrain	
			dataset 4 (3 months)	dorsal patterned forebrain	
			dataset 5 (3 months)	dorsal patterned forebrain	
			dataset 7 (3 months)	dorsal patterned forebrain	
			Mito210c1 (4 months)	dorsal patterned forebrain	
			PGP1 (4 months)	dorsal patterned forebrain	
			Mito210c1 (5 months)	dorsal patterned forebrain	
			PGP1 (5 months)	dorsal patterned forebrain	
			dataset 1 (6 months)	dorsal patterned forebrain	
			dataset 2 (6 months)	dorsal patterned forebrain	
			dataset 3 (6 months)	dorsal patterned forebrain	
			dataset 4 (6 months)	dorsal patterned forebrain	
			dataset 5 (6 months)	dorsal patterned forebrain	
			dataset 7 (6 months)	dorsal patterned forebrain	
	Fiddes et al., Field et al., Sanders et al.	Cell line/Age	GSM2867931	cortical	GEO: GSE106245
			GSM2867932	cortical	
			GSM2867933	cortical	
			GSM2867934	cortical	
			GSM2867935	cortical	
			GSM2867936	cortical	
			GSM2867937	cortical	
	Xiang et al.	NA	NA	thalamus	GEO: GSE122342
	Khan et al.	Control/Patient	GSM4306931	cortical	GEO: GSE145122
			GSM4306932	cortical	
			GSM4306933	cortical	
			GSM4306934	cortical	
	Nayler et al.	Matrigel encapsulated	GSM4524697	cerebellum	GEO: GSE150153

		unencapsulated	GSM4524699	cerebellum	
	Fair et al.	Age	GSM4750931	cerebral	GEO: GSE157019
			GSM4750932	cerebral	
	Parisian et al.	Control/SMARCB1 knockdown	GSM4769380	cerebral	GEO: GSE157525
			GSM4769381	cerebral	
			GSM4769382	cerebral	
			GSM4769383	cerebral	
			GSM4769384	cerebral	
			GSM4769385	cerebral	
	Chen et al.	Control/Serum exposed	GSM4996460	cerebral	GEO: GSE164089
			GSM4996461	cerebral	
			GSM4996462	cerebral	
			GSM4996463	cerebral	
	Dailamy et al.	Induction/No induction	GSM5005486	neural_induced_blood_vessel	GEO: GSE164268
			GSM5005487	neural_induced_blood_vessel	
			GSM5005488	neural_induced_blood_vessel	
	Banfi et al.	Control/SETBP1 mutant	GSM5221533	cerebral	GEO: GSE171263
			GSM5221534	cerebral	
	Popova et al.	Batch	GSM5478754	cortical	GEO: GSE180945
			GSM5478755	cortical	
			GSM5478756	cortical	
	Suong et al.	orbital/vertical mixing	GSM5587100	cerebral	GEO: GSE184409
			GSM5587101	cerebral	
			GSM5587102	cerebral	
			GSM5587103	cerebral	
			GSM5587104	cerebral	
			GSM5587105	cerebral	
	Xiang and Tanaka et al.	NA	NA	MGE_and_cortical	GEO: GSE98201
	Kronenberg et al.	NA	NA	cerebral	GEO: GSE113931
	Pollen et al.	NA	NA	cerebral	GEO: GSE124299
	Velasco et al.	Cell line/age	GSE129519_expression_11a.6mon.txt	dorsal patterned forebrain	GEO: GSE129519
			GSE129519_expression_HUES66.3mon.txt	dorsal patterned forebrain	
			GSE129519_expression_PGP1.3mon.txt	dorsal patterned forebrain	
			GSE129519_expression_PGP1.6mon.txt	dorsal patterned forebrain	

			GSE129519_expression_GM.6mon.txt	dorsal patterned forebrain	
			GSE129519_expression_PGP1.3mon.batch2.txt	dorsal patterned forebrain	
			GSE129519_expression_PGP1.6mon.batch3.txt	dorsal patterned forebrain	
	Shi et al.	NA	NA	vascularized_cortical	GEO: GSE131094
	Qian et al.	NA	NA	cortical	GEO: GSE137941
	Eura et al.	NA	NA	brainstem	GEO: GSE145306
	Huang et al.	Age/Batch	GSM4996689	hypothalamic_arcuate	GEO: GSE164102
			GSM4996690	hypothalamic_arcuate	
			GSM4996691	hypothalamic_arcuate	
			GSM4996692	hypothalamic_arcuate	
			GSM4996693	hypothalamic_arcuate	
			GSM4996694	hypothalamic_arcuate	
			GSM4996695	hypothalamic_arcuate	
			GSM4996696	hypothalamic_arcuate	
			GSM4996697	hypothalamic_arcuate	
			GSM4996698	hypothalamic_arcuate	
			GSM4996699	hypothalamic_arcuate	
			GSM4996700	hypothalamic_arcuate	
			GSM4996701	hypothalamic_arcuate	
	Fiorenzano et al.	Age	GSM5136255	ventral_midbrain	GEO: GSE168323
			GSM5136256	ventral_midbrain	
			GSM5136257	ventral_midbrain	
			GSM5136258	ventral_midbrain	
			GSM5136259	ventral_midbrain	
			GSM5136260	ventral_midbrain	
			GSM5136261	ventral_midbrain	
			GSM5136262	ventral_midbrain	
			GSM5136263	ventral_midbrain	
			GSM5136264	ventral_midbrain	
			GSM5136265	ventral_midbrain	
			GSM5136266	ventral_midbrain	
			GSM5136267	ventral_midbrain	
			GSM5136268	ventral_midbrain	
			GSM5136269	ventral_midbrain	
			GSM5136270	ventral_midbrain	
			GSM5136271	ventral_midbrain	

			GSM5136272	ventral_midbrain	
			GSM5136273	ventral_midbrain	
			GSM5136274	ventral_midbrain	
			GSM5136275	ventral_midbrain	
			GSM5136276	ventral_midbrain	
			GSM5136277	ventral_midbrain	
			GSM5136278	ventral_midbrain	
			GSM5136279	ventral_midbrain	
			GSM5136280	ventral_midbrain	
			GSM5136281	ventral_midbrain	
	Fernando et al.	NA	NA	cortical_and_neural_retina	GEO: GSE174232
	Szebényi et al.	Batch	GSE180122_C9_batch1_data.csv	cortical	GEO: GSE180122
			GSE180122_C9_batch2_data.csv	cortical	
	Quadrato et al.	Age	GSM2295945	cerebral	GEO: GSE86153
			GSM2295946	cerebral	
	Revah et al.	Age/Transplanted	GSM5732392	cortical	GEO: GSE190815
			GSM5732393	cortical	
			GSM5732394	cortical	
			GSM6225773	cortical	
			GSM6225774	cortical	
			GSM6225775	cortical	

Supplemental Table S4.2

Class-level grouping of author provided cell-type annotations

Table containing our mapping between author provided annotations (Author annotations column) and our broad cell-type annotations (Class annotations column).

System	**Study**	**Author annotations**	**Class annotations**
organoid	Uzquiano et al.	aRG	Neural Progenitor
		oRG	Neural Progenitor
		oRG II	Neural Progenitor
		Subcortical progenitors	Neural Progenitor
		IN progenitors	Neural Progenitor
		IP	Intermediate Progenitor
			Dividing Progenitor
		CFuPN	Glutamatergic
		CPN	Glutamatergic
		FOXG1-EMX1-neurons	Glutamatergic
		Newborn CFuPN	Glutamatergic
		Newborn CPN	Glutamatergic
		Newborn DL PN	Glutamatergic
		Newborn PN	Glutamatergic
		PN	Glutamatergic
		Subcortical neuronal precursors	Glutamatergic
		Subcoritcal neurons	Glutamatergic
		Preplate/Subplate	Glutamatergic
		Immature IN	GABAergic
		Subcortical interneurons	GABAergic
		Astroglia	Non-neuronal
		oRG/Astroglia	Non-neuronal
		Glial precursors	Non-neuronal
		Mesenchyme	Non-neuronal
		Unknown	other
		Cajal Retzius	other
		Cortical hem	other
		Neural crest	other
		Neural placode	other

		Pre-delaminating neural crest	other
annotated fetal	Polioudakis et. al.	vRG	Neural Progenitor
		oRG	Neural Progenitor
		OPC	Neural Progenitor
		IP	Intermediate Progenitor
		PgG2M	Dividing Progenitor
		PgS	Dividing Progenitor
		ExDp1	Glutamatergic
		ExDp2	Glutamatergic
		ExM	Glutamatergic
		ExM-U	Glutamatergic
		ExN	Glutamatergic
		InMGE	GABAergic
		InCGE	GABAergic
		End	Non-neuronal
		Mic	Non-neuronal
		Per	Non-neuronal
	Fan et al.	NPC	Neural Progenitor
			Dividing Progenitor
		EX_cor	Glutamatergic
		IN_cor	GABAergic
		Endo	Non-neuronal
		Astro	Non-neuronal
		Blood	Non-neuronal
		Immune	Non-neuronal
		Oligo	Non-neuronal
		CR	other
		PONS_neu	other
	Bhaduri et al.	RG	Neural Progenitor
		IPC	Intermediate Progenitor
		Dividing	Dividing Progenitor
		Neuron	Glutamatergic
		Interneuron	GABAergic
		Astrocyte	Non-neuronal
		Endo	Non-neuronal
		Microglia	Non-neuronal
		Vascular	Non-neuronal

		Oligo	Non-neuronal
	Braun et al.		Dividing Progenitor
		RGL | O-HEM	Neural Progenitor
		RGL | GBL | O-HEM | S-CC	Neural Progenitor
		RGL | GBL | O-HEM | P-FPL	Neural Progenitor
		RGL | GBL | O-HEM	Neural Progenitor
		RGL | GBL | S-CC | S-G1S	Neural Progenitor
		RGL | GBL | O-HEM | S-CC | S-G1S	Neural Progenitor
		RGL | GBL | S-CC	Neural Progenitor
		RGL | GBL	Neural Progenitor
		RGL | P-FPL	Neural Progenitor
		RGL	Neural Progenitor
		RGL | GBL | P-FPL	Neural Progenitor
		RGL | O-HEM | P-FPL	Neural Progenitor
		RGL | P-FP1 | P-FPL	Neural Progenitor
		RGL | P-FP1	Neural Progenitor
		RGL | S-CC	Neural Progenitor
		RGL | S-CC | S-G1S	Neural Progenitor
		RGL | O-COP | S-CC | S-G1S	Neural Progenitor
		RGL | O-COP | O-HEM | S-CC | S-G1S	Neural Progenitor
		RGL | O-HEM | S-CC | S-G1S	Neural Progenitor
		RGL | M-CHRP | O-HEM	Neural Progenitor
		NBL | NEUR | RGL | S-CC | S-G1S	Neural Progenitor
		NEUR | RGL | S-CC | S-G1S	Neural Progenitor
		NBL | RGL	Neural Progenitor
		RGL | M-PER | S-CC | S-G1S	Neural Progenitor
		RGL | M-PER	Neural Progenitor
		RGL | P-TEL | S-CC	Neural Progenitor
		RGL | P-PALL-M | P-PALL | P-TEL | S-CC | S-G1S	Neural Progenitor
		RGL | GBL | P-PALL | P-TEL | S-CC | S-G1S	Neural Progenitor
		RGL | P-PALL | P-TEL | S-CC | S-G1S	Neural Progenitor
		RGL | P-PALL | P-TEL	Neural Progenitor
		RGL | GBL | P-PALL-M | P-PALL | P-TEL | S-CC | S-G1S	Neural Progenitor
		NBL | RGL | S-CC	Neural Progenitor

		NBL | RGL | P-PALL-M | P-PALL	Neural Progenitor
		NBL | RGL | P-PALL-M | P-PALL | P-TEL	Neural Progenitor
		NBL | S-CC | S-G1S	Intermediate Progenitor
		NBL | NEUR | S-CC | S-G2M	Intermediate Progenitor
		NBL | NEUR | S-CC | S-G1S	Intermediate Progenitor
		NBL | NEUR | NT-GABA | S-CC | S-G1S	Intermediate Progenitor
		S-CC | S-G1S | S-G2M	Intermediate Progenitor
		NBL | NEUR | M-PER | S-CC | S-G2M	Intermediate Progenitor
		NEUR | NT-GABA | P-VLGE | P-SUBPALL | P-TEL | S-CC | S-G2M	Intermediate Progenitor
		NT-GABA | P-VLGE | P-SUBPALL | P-TEL | S-CC | S-G2M	Intermediate Progenitor
		NEUR | NT-GABA | P-VLGE | P-SUBPALL | P-TEL | S-CC | S-G1S	Intermediate Progenitor
		NEUR | NT-GABA | P-DLGE | P-SUBPALL | P-TEL | S-CC | S-G2M	Intermediate Progenitor
		NEUR | NT-GABA | P-DLGE | P-SUBPALL | P-TEL | S-CC | S-G1S	Intermediate Progenitor
		NEUR | NT-GABA | P-DLGE | P-SUBPALL | P-TEL	Intermediate Progenitor
		NBL | NEUR | S-CC | S-G1S | S-G2M	Intermediate Progenitor
		NBL | NEUR | S-CC	Intermediate Progenitor
		NBL	Intermediate Progenitor
		NBL | NEUR | NT-VGLUT2 | P-PALL | P-TEL	Intermediate Progenitor
		NBL | NEUR | P-PALL-M | P-PALL | P-TEL | S-CC | S-G1S | S-G2M	Intermediate Progenitor
		NBL | P-PALL-M | P-PALL | P-TEL	Intermediate Progenitor
		NBL | P-PALL | P-TEL | S-CC | S-G1S	Intermediate Progenitor
		NBL | NEUR | P-PALL | P-TEL | S-CC | S-G1S | S-G2M	Intermediate Progenitor
		NBL | NEUR | NT-VGLUT2 | P-PALL | P-TEL | S-CC | S-G2M	Intermediate Progenitor
		TH-RETN | NEUR | NT-GABA | P-DLGE | P-SUBPALL | P-TEL | S-CC | S-G2M	Intermediate Progenitor
		TH-RETN | NEUR | RGL | NT-GABA | P-DLGE | P-SUBPALL | P-TEL | S-CC | S-G1S	Intermediate Progenitor
		NBL | NEUR | M-PER	Intermediate Progenitor
		NBL | NEUR	Intermediate Progenitor
		NEUR | S-CC | S-G1S | S-G2M	Intermediate Progenitor
		RGL | GBL | P-TEL	Glutamatergic
		RGL | GBL | NT-VGLUT3	Glutamatergic
		RGL | GBL | M-CHRP | O-HEM | P-PALL-M | P-PALL	Glutamatergic
		RGL | O-HEM | P-PALL-M | P-PALL | S-CC | S-G1S	Glutamatergic
		RGL | O-COP | P-TEL | S-CC | S-G1S	Glutamatergic
		RGL | GBL | O-HEM | P-PALL-M | P-PALL	Glutamatergic
		RGL | P-PALL-M | P-PALL | S-CC | S-G1S	Glutamatergic

		RGL | P-TEL | S-CC | S-G1S	Glutamatergic
		RGL | NT-VGLUT3 | S-CC	Glutamatergic
		NBL | NEUR | NP-TRH | NT-VGLUT2 | P-PALL-M | P-PALL | S-CC	Glutamatergic
		NBL | NEUR | NT-VGLUT2 | P-PALL-M | P-PALL | P-TEL	Glutamatergic
		NBL | NEUR | NT-VGLUT1 | NT-VGLUT2 | P-PALL | P-TEL	Glutamatergic
		NEUR | NT-VGLUT1 | NT-VGLUT2 | P-PALL | P-TEL	Glutamatergic
		NEUR | NT-VGLUT1 | P-PALL | P-TEL	Glutamatergic
		NBL | NEUR | RGL | NT-VGLUT1 | NT-VGLUT2 | P-PALL-M | P-PALL | P-TEL | S-CC | S-G1S	Glutamatergic
		NBL | NEUR | NT-VGLUT1 | NT-VGLUT2 | P-PALL-M | P-PALL | P-TEL	Glutamatergic
		NBL | NEUR | NT-VGLUT1 | P-PALL | P-TEL	Glutamatergic
		NEUR | NT-VGLUT2	Glutamatergic
		NBL | NEUR | NT-VGLUT2	Glutamatergic
		NEUR | NP-TRH | NT-VGLUT2	Glutamatergic
		NEUR | RGL | NP-TRH | NT-VGLUT2	Glutamatergic
		NEUR | NT-SER | NT-VGLUT3	Glutamatergic
		NEUR | NT-VGLUT3	Glutamatergic
		NEUR | NT-VGLUT2 | NT-VGLUT3	Glutamatergic
		NBL | NEUR | NP-TRH | NT-VGLUT2 | P-TEL	Glutamatergic
		NBL | NEUR | NP-TRH | NT-VGLUT2 | P-PALL-M | P-PALL	Glutamatergic
		NBL | NEUR | NP-TRH | NT-VGLUT2 | O-HEM | P-PALL-M | P-PALL	Glutamatergic
		NEUR | NT-VGLUT2 | P-TEL	Glutamatergic
		NEUR | NP-TRH | NT-VGLUT2 | P-TEL	Glutamatergic
		NBL | NEUR | NT-VGLUT2 | P-TEL	Glutamatergic
		NBL | NEUR | NP-AVP | NP-POMC | NP-TRH | NT-VGLUT2 | P-TEL	Glutamatergic
		NEUR | NP-POMC | NT-VGLUT2	Glutamatergic
		NEUR | NP-HCRT | NT-VGLUT3	Glutamatergic
		NBL | NEUR | M-PER | NT-VGLUT2	Glutamatergic
		NBL | NEUR | M-PER | NP-TRH | NT-VGLUT2	Glutamatergic
		NBL | NEUR | M-PER | NP-HCRT | NT-VGLUT2	Glutamatergic
		NBL | NEUR | NP-HCRT | NT-VGLUT2	Glutamatergic
		NBL | NEUR | NP-TRH | NT-VGLUT2	Glutamatergic
		HB-OTV | NBL | NEUR | NT-VGLUT1 | NT-VGLUT2	Glutamatergic
		HB-OTV | NBL | NEUR | GBL | NT-VGLUT1	Glutamatergic
		HB-OTV | NBL | NEUR | NT-VGLUT2 | P-TEL	Glutamatergic
		HB-OTV | NBL | NEUR | NT-VGLUT1 | NT-VGLUT2 | P-TEL	Glutamatergic
		HB-OTV | NEUR | RGL | NT-VGLUT1 | NT-VGLUT2 | P-TEL	Glutamatergic
		HB-OTV | NBL | NEUR | NP-POMC | NT-VGLUT1 | NT-VGLUT2	Glutamatergic
		NBL | NEUR | NP-TRH | NT-VGLUT1 | NT-VGLUT2	Glutamatergic

		NEUR | RGL | NT-GABA | O-HEM | P-DLGE | P-PALL-M | P-PALL | P-SUBPALL | P-TEL | S-CC | S-G1S	GABAergic
		NBL | NEUR | NT-GABA	GABAergic
		RGL | P-DLGE | P-PALL-M | P-PALL | P-SUBPALL | P-TEL | S-CC | S-G1S	GABAergic
		RGL | GBL | P-DLGE | P-SUBPALL | P-TEL | S-CC | S-G1S	GABAergic
		RGL | NT-GABA | P-PALL | P-TEL | S-CC | S-G1S	GABAergic
		RGL | GBL | P-VLGE | P-SUBPALL | P-TEL | S-CC | S-G1S	GABAergic
		RGL | GBL | NT-GABA | P-DLGE | P-SUBPALL | P-TEL | S-CC | S-G1S	GABAergic
		NEUR | NT-GABA | P-VLGE | P-SUBPALL | P-TEL	GABAergic
		NT-GABA | P-SUBPALL | P-TEL | S-CC	GABAergic
		TH-RETN | P-VLGE | P-SUBPALL | P-TEL	GABAergic
		TH-RETN | NEUR | NT-GABA | P-VLGE | P-SUBPALL | P-TEL	GABAergic
		TH-RETN | NEUR | NT-GABA | P-DLGE	GABAergic
		NEUR | NT-GABA	GABAergic
		TH-RETN | NEUR | NT-GABA | P-VLGE	GABAergic
		TH-RETN | NEUR | NT-GABA	GABAergic
		NEUR | NT-GABA | P-TEL	GABAergic
		NEUR | NT-GABA | P-SUBPALL | P-TEL	GABAergic
		TH-RETN | NEUR | NT-GABA | P-DLGE | P-SUBPALL | P-TEL	GABAergic
		NEUR | NP-GNRH | NT-GABA | P-TEL	GABAergic
		NEUR | NP-GNRH | NT-GABA | P-DLGE | P-SUBPALL | P-TEL	GABAergic
		TH-RETN | NEUR | NP-GNRH | NT-GABA | P-VLGE | P-SUBPALL | P-TEL	GABAergic
		NEUR | NP-TRH | NT-GABA	GABAergic
		NEUR | NT-GABA | P-PALL	GABAergic
		NEUR | NT-GABA | P-DLGE	GABAergic
		NEUR | NT-GABA | P-VLGE	GABAergic
		TH-RETN | NEUR | NP-GNRH | NT-GABA | P-DLGE | P-SUBPALL | P-TEL	GABAergic
		NEUR | NP-GNRH | NT-GABA | P-VLGE | P-SUBPALL | P-TEL	GABAergic
		NEUR | M-PER | NP-POMC | NT-GABA	GABAergic
		NEUR | M-PER | NT-GABA | P-VLGE	GABAergic
		NEUR | M-PER | NT-GABA	GABAergic
		NBL | NEUR | M-PER | NT-GABA	GABAergic
		RGL | M-CHRP | M-PER | O-HEM | S-CC | S-G1S	Non-neuronal
		M-ERY	Non-neuronal
		M-ERY | NP-POMC | S-CC	Non-neuronal
		M-IMMUNE | M-MGL | M-PVM | S-CC	Non-neuronal
		M-IMMUNE | M-MGL | M-PVM	Non-neuronal
		RGL | M-IMMUNE | M-MGL | M-PVM | NP-POMC | S-CC	Non-neuronal

		M-IMMUNE | M-MGL	Non-neuronal
		M-IMMUNE	Non-neuronal
		RGL | M-PER | M-ENDO | NP-TRH	Non-neuronal
		RGL | M-ENDO	Non-neuronal
		RGL | M-PER | M-ENDO | NP-TRH | S-CC	Non-neuronal
		RGL | M-ENDO | NP-TRH	Non-neuronal
		RGL | M-PER | M-FBL	Non-neuronal
		RGL | M-PER | M-FBL | M-VSMC	Non-neuronal
		RGL | M-MGL | M-PER | M-FBL | S-CC | S-G1S	Non-neuronal
		RGL | M-PER | M-FBL | NP-POMC | P-FP1 | S-CC | S-G1S	Non-neuronal
		RGL | M-FBL | P-VLGE | S-CC | S-G1S	Non-neuronal
		RGL | M-PER | M-FBL | P-VLGE | S-CC | S-G1S	Non-neuronal
		RGL | M-FBL | S-CC | S-G1S	Non-neuronal
		RGL | M-FBL | NP-POMC | S-CC | S-G1S	Non-neuronal
		RGL | M-PER | M-FBL | NP-POMC | S-CC	Non-neuronal
		RGL | M-PER | M-FBL | NP-POMC | P-FP1 | S-CC	Non-neuronal
		RGL | M-PER | M-FBL | S-CC	Non-neuronal
		RGL | M-FBL | NP-TRH	Non-neuronal
		RGL | M-FBL | P-FP1	Non-neuronal
		RGL | M-PER | M-FBL | NP-POMC | P-VLGE | S-CC | S-G1S	Non-neuronal
		E-SCHWL | RGL | NP-POMC | P-VLGE | S-CC | S-G1S	Non-neuronal
		E-SCHWL | RGL | M-FBL	Non-neuronal
		NEUR | OPC	Non-neuronal
		OPC	Non-neuronal
		RGL | OPC	Non-neuronal
		NEUR | RGL | OPC	Non-neuronal
		RGL | OPC | P-TEL	Non-neuronal
		NBL | NEUR | RGL | GBL | S-CC | S-G1S	other
		NBL | RGL | S-CC | S-G1S	other
		RGL | GBL | P-TEL | S-CC | S-G1S	other
			other
		NBL | NEUR | NT-VGLUT1 | NT-VGLUT2 | NT-GABA | P-PALL | P-VLGE | P-SUBPALL | P-TEL	other
		TH-RETN | NEUR | NT-VGLUT2 | NT-GABA | P-DLGE | P-SUBPALL | P-TEL	other
		NEUR	other
		HB-OTV | NEUR	other
		NEUR | NT-SER	other
		NEUR | NT-GLY	other
		NBL | NEUR | NT-GLY	other

		NEUR | NT-VGLUT3 | NT-GABA | P-FP1	other
		NBL | NEUR | M-PER | NT-GLY	other
		HB-OTV | NBL | P-TEL | S-CC	other
		NBL | NEUR | NP-TRH	other

8. Appendix 2 – References for Supplemental Table 4.1.

Banfi F, Rubio A, Zaghi M, Massimino L, Fagnocchi G, Bellini E, Luoni M, Cancellieri C, Bagliani A, Di Resta C, et al. 2021. SETBP1 accumulation induces P53 inhibition and genotoxic stress in neural progenitors underlying neurodegeneration in Schinzel-Giedion syndrome. *Nat Commun* **12**: 4050.

Bhaduri A, Sandoval-Espinosa C, Otero-Garcia M, Oh I, Yin R, Eze UC, Nowakowski TJ, Kriegstein AR. 2021. An atlas of cortical arealization identifies dynamic molecular signatures. *Nature* **598**: 200–204.

Braun E, Danan-Gotthold M, Borm LE, Vinsland E, Lee KW, Lönnerberg P, Hu L, Li X, He X, Andrusivová Ž, et al. 2022. Comprehensive cell atlas of the first-trimester developing human brain. 2022.10.24.513487. https://www.biorxiv.org/content/10.1101/2022.10.24.513487v1 (Accessed October 31, 2022).

Chen X, Sun G, Tian E, Zhang M, Davtyan H, Beach TG, Reiman EM, Blurton-Jones M, Holtzman DM, Shi Y. 2021. Modeling Sporadic Alzheimer’s Disease in Human Brain Organoids under Serum Exposure. *Adv Sci* **8**: 2101462.

Dailamy A, Parekh U, Katrekar D, Kumar A, McDonald D, Moreno A, Bagheri P, Ng TN, Mali P. 2021. Programmatic introduction of parenchymal cell types into blood vessel organoids. *Stem Cell Rep* **16**: 2432–2441.

Eura N, Matsui TK, Luginbühl J, Matsubayashi M, Nanaura H, Shiota T, Kinugawa K, Iguchi N, Kiriyama T, Zheng C, et al. 2020. Brainstem Organoids From Human Pluripotent Stem Cells. *Front Neurosci* **14**. https://www.frontiersin.org/articles/10.3389/fnins.2020.00538 (Accessed January 26, 2023).

Fair SR, Julian D, Hartlaub AM, Pusuluri ST, Malik G, Summerfied TL, Zhao G, Hester AB, Ackerman WE, Hollingsworth EW, et al. 2020. Electrophysiological Maturation of Cerebral Organoids Correlates with Dynamic Morphological and Cellular Development. *Stem Cell Rep* **15**: 855–868.

Fan X, Fu Y, Zhou X, Sun L, Yang M, Wang M, Chen R, Wu Q, Yong J, Dong J, et al. 2020. Single-cell transcriptome analysis reveals cell lineage specification in temporal-spatial patterns in human cortical development. *Sci Adv* **6**: eaaz2978.

Fernando M, Lee S, Wark JR, Xiao D, Lim BY, O’Hara-Wright M, Kim HJ, Smith GC, Wong T, Teber ET, et al. 2022. Differentiation of brain and retinal organoids from confluent cultures of pluripotent stem cells connected by nerve-like axonal projections of optic origin. *Stem Cell Rep* **17**: 1476–1492.

Fiddes IT, Lodewijk GA, Mooring M, Bosworth CM, Ewing AD, Mantalas GL, Novak AM, van den Bout A, Bishara A, Rosenkrantz JL, et al. 2018. Human-Specific NOTCH2NL Genes Affect Notch Signaling and Cortical Neurogenesis. *Cell* **173**: 1356-1369.e22.

Field AR, Jacobs FMJ, Fiddes IT, Phillips APR, Reyes-Ortiz AM, LaMontagne E, Whitehead L, Meng V, Rosenkrantz JL, Olsen M, et al. 2019. Structurally Conserved Primate LncRNAs Are Transiently Expressed during Human Cortical Differentiation and Influence Cell-Type-Specific Genes. *Stem Cell Rep* **12**: 245–257.

Fiorenzano A, Sozzi E, Birtele M, Kajtez J, Giacomoni J, Nilsson F, Bruzelius A, Sharma Y, Zhang Y, Mattsson B, et al. 2021. Single-cell transcriptomics captures features of human midbrain development and dopamine neuron diversity in brain organoids. *Nat Commun* **12**: 7302.

Huang W-K, Wong SZH, Pather SR, Nguyen PTT, Zhang F, Zhang DY, Zhang Z, Lu L, Fang W, Chen L, et al. 2021. Generation of hypothalamic arcuate organoids from human induced pluripotent stem cells. *Cell Stem Cell* **28**: 1657-1670.e10.

Jorstad NL, Song JHT, Exposito-Alonso D, Suresh H, Castro N, Krienen FM, Yanny AM, Close J, Gelfand E, Travaglini KJ, et al. 2022. *Comparative transcriptomics reveals human-specific cortical features*. Neuroscience http://biorxiv.org/lookup/doi/10.1101/2022.09.19.508480 (Accessed January 11, 2023).

Khan TA, Revah O, Gordon A, Yoon S-J, Krawisz AK, Goold C, Sun Y, Kim CH, Tian Y, Li M-Y, et al. 2020. Neuronal defects in a human cellular model of 22q11.2 deletion syndrome. *Nat Med* **26**: 1888–1898.

Kronenberg ZN, Fiddes IT, Gordon D, Murali S, Cantsilieris S, Meyerson OS, Underwood JG, Nelson BJ, Chaisson MJP, Dougherty ML, et al. 2018. High-resolution comparative analysis of great ape genomes. *Science* **360**: eaar6343.

Nayler S, Agarwal D, Curion F, Bowden R, Becker EBE. 2021. High-resolution transcriptional landscape of xeno-free human induced pluripotent stem cell-derived cerebellar organoids. *Sci Rep* **11**: 12959.

Parisian AD, Koga T, Miki S, Johann PD, Kool M, Crawford JR, Furnari FB. 2020. SMARCB1 loss interacts with neuronal differentiation state to block maturation and impact cell stability. *Genes Dev* **34**: 1316–1329.

Polioudakis D, de la Torre-Ubieta L, Langerman J, Elkins AG, Shi X, Stein JL, Vuong CK, Nichterwitz S, Gevorgian M, Opland CK, et al. 2019. A Single-Cell Transcriptomic Atlas of Human Neocortical Development during Mid-gestation. *Neuron* **103**: 785-801.e8.

Pollen AA, Bhaduri A, Andrews MG, Nowakowski TJ, Meyerson OS, Mostajo-Radji MA, Di Lullo E, Alvarado B, Bedolli M, Dougherty ML, et al. 2019. Establishing Cerebral Organoids as Models of Human-Specific Brain Evolution. *Cell* **176**: 743-756.e17.

Popova G, Soliman SS, Kim CN, Keefe MG, Hennick KM, Jain S, Li T, Tejera D, Shin D, Chhun BB, et al. 2021. Human microglia states are conserved across experimental models and regulate neural stem cell responses in chimeric organoids. *Cell Stem Cell* **28**: 2153-2166.e6.

Qian X, Su Y, Adam CD, Deutschmann AU, Pather SR, Goldberg EM, Su K, Li S, Lu L, Jacob F, et al. 2020. Sliced Human Cortical Organoids for Modeling Distinct Cortical Layer Formation. *Cell Stem Cell* **26**: 766-781.e9.

Quadrato G, Nguyen T, Macosko EZ, Sherwood JL, Min Yang S, Berger DR, Maria N, Scholvin J, Goldman M, Kinney JP, et al. 2017. Cell diversity and network dynamics in photosensitive human brain organoids. *Nature* **545**: 48–53.

Revah O, Gore F, Kelley KW, Andersen J, Sakai N, Chen X, Li M-Y, Birey F, Yang X, Saw NL, et al. 2022. Maturation and circuit integration of transplanted human cortical organoids. *Nature* **610**: 319–326.

Shi Y, Sun L, Wang M, Liu J, Zhong S, Li R, Li P, Guo L, Fang A, Chen R, et al. 2020. Vascularized human cortical organoids (vOrganoids) model cortical development in vivo. *PLOS Biol* **18**: e3000705.

Shi Y, Wang M, Mi D, Lu T, Wang B, Dong H, Zhong S, Chen Y, Sun L, Zhou X, et al. 2021. Mouse and human share conserved transcriptional programs for interneuron development. *Science* **374**: eabj6641.

Suong DNA, Imamura K, Inoue I, Kabai R, Sakamoto S, Okumura T, Kato Y, Kondo T, Yada Y, Klein WL, et al. 2021. Induction of inverted morphology in brain organoids by vertical-mixing bioreactors. *Commun Biol* **4**: 1–13.

Szebényi K, Wenger LMD, Sun Y, Dunn AWE, Limegrover CA, Gibbons GM, Conci E, Paulsen O, Mierau SB, Balmus G, et al. 2021. Human ALS/FTD brain organoid slice cultures display distinct early astrocyte and targetable neuronal pathology. *Nat Neurosci* **24**: 1542–1554.

Trevino AE, Müller F, Andersen J, Sundaram L, Kathiria A, Shcherbina A, Farh K, Chang HY, Pașca AM, Kundaje A, et al. 2021. Chromatin and gene-regulatory dynamics of the developing human cerebral cortex at single-cell resolution. *Cell* **184**: 5053-5069.e23.

Uzquiano A, Kedaigle AJ, Pigoni M, Paulsen B, Adiconis X, Kim K, Faits T, Nagaraja S, Antón-Bolaños N, Gerhardinger C, et al. 2022. Proper acquisition of cell class identity in organoids allows definition of fate specification programs of the human cerebral cortex. *Cell* **185**: 3770-3788.e27.

Velasco S, Kedaigle AJ, Simmons SK, Nash A, Rocha M, Quadrato G, Paulsen B, Nguyen L, Adiconis X, Regev A, et al. 2019. Individual brain organoids reproducibly form cell diversity of the human cerebral cortex. *Nature* **570**: 523–527.

Xiang Y, Tanaka Y, Cakir B, Patterson B, Kim K-Y, Sun P, Kang Y-J, Zhong M, Liu X, Patra P, et al. 2019. hESC-Derived Thalamic Organoids Form Reciprocal Projections When Fused with Cortical Organoids. *Cell Stem Cell* **24**: 487-497.e7.

Xiang Y, Tanaka Y, Patterson B, Kang Y-J, Govindaiah G, Roselaar N, Cakir B, Kim K-Y, Lombroso AP, Hwang S-M, et al. 2017. Fusion of regionally-specified hPSC-derived organoids models human brain development and interneuron migration. *Cell Stem Cell* **21**: 383-398.e7.

Yu Y, Zeng Z, Xie D, Chen R, Sha Y, Huang S, Cai W, Chen W, Li W, Ke R, et al. 2021. Interneuron origin and molecular diversity in the human fetal brain. *Nat Neurosci* **24**: 1745–1756.

Zhou X, Lu Y, Zhao F, Dong J, Ma W, Zhong S, Wang M, Wang B, Zhao Y, Shi Y, et al. 2022. Deciphering the spatial-temporal transcriptional landscape of human hypothalamus development. *Cell Stem Cell* **29**: 328-343.e5.