EPIGENETICS AND GENETICS
OF HEMATOPOIETIC STEM CELLS HETEROGENEITY

NADIYA KUBASOVA

Tese para obtenção do grau de Doutor em Envelhecimento e Doenças Crónicas

Doutoramento em associação entre:

Universidade NOVA de Lisboa (Faculdade de Ciências Médicas | NOVA Medical School - FCM|NMS/UNL)

Universidade de Coimbra (Faculdade de Medicina - FM/UC)

Universidade do Minho (Escola de Medicina - EMed/UM)

Fevereiro, 2022

EPIGENETICS AND GENETICS OF HEMATOPOIETIC STEM CELLS HETEROGENEITY

Nadiya Kubasova
Vasco M. Barreto, Investigador Principal da FCM | NMS/UNL
António Gil Castro, Professor Associado da EMed/UM

Tese para obtenção do grau de Doutor em Envelhecimento e Doenças Crónicas

Tese para obtenção do grau de Doutor em Envelhecimento e Doenças Crónicas
Doutoramento em associação entre:
Universidade NOVA de Lisboa (Faculdade de Ciências Médicas | NOVA Medical School - FCM|NMS/UNL)
Universidade de Coimbra (Faculdade de Medicina - FM/UC)
Universidade do Minho (Escola de Medicina - EMed/UM)

PD + F PROGRAMAS DE DOUTORAMENTO FCT

Fevereiro, 2022

Preface

This work was developed at the Chronic Diseases Research Centre (CEDOC) of Faculdade de Ciências Médicas (FCM) | Nova Medical School (NMS). It was supported by the FCT (Fundação para a Ciência e a Tecnologia) with grants PTDC/BEX-BCM/5900/2014 and IF/01721/2014/CP1252/CT0005. and with a Ph.D. fellowship PD/BD/114164/2016 attributed by the Inter-University Doctoral Programme in Ageing and Chronic Diseases (PhDOC). Ph.D. fellowship was attributed by FCT, from FSE (Fundo Social Europeu) and POCH (Programa Operacional Capital Humano).

Results obtained during this dissertation produced

publication of papers in peer-reviewed journals:

Research article

Kubasova, N., Alves-Pereira, C.F., Gupta, S., Vinogradova, S., Gimelbrant, A., and Barreto, V.M. (2022). In vivo clonal analysis reveals random monoallelic expression in lymphocytes that traces back to hematopoietic stem cells. Front. Cell Dev. Biol. *in press*. doi: 10.3389/fcell.2022.827774.

Review article

Barreto, V.M., Kubasova, N., Alves-Pereira, C.F., and Gendrel, A.-V. (2021). X-Chromosome Inactivation and Autosomal Random Monoallelic Expression as "Faux Amis". Front. Cell Dev. Biol. *9*, 2599. doi: 10.3389/fcell.2021.740937.

reposition in NCBI Gene Expression Omnibus:

RNA-seq and WES-seq expression data under series accession number GEO: GSE174040.

The procedures in this thesis involving animals have been approved by Órgão Responsável pelo Bem-Estar dos Animais (ORBEA) of Instituto Gulbenkian de Ciência (PTDC/BEX-BCM/5900/2014 reference).

Acknowledgments

All scientific work is the outcome of more than one person's contribution, and this thesis is no exception; it is the product of the collective effort of people that surrounded me over these years and without whom it would not have been possible to complete this work.

In the first place, I would like to thank my supervisors Vasco M. Barreto and António Gil Castro for accepting me as a Ph.D. student, with a special acknowledgment to Vasco for receiving me in his team, allowing me to develop this project under his guidance, and for his confidence. It was a special experience to develop this Ph.D. project with him over these years. He gave me support, advice, and motivation every time I needed it, and helped me at the bench with the reconstitution experiments that took long hours.

I am also deeply grateful to Alexander Gimelbrant for his kind collaboration, impressive experience, and crucial contribution for the results. I thank him for believing in this project and the endless patience during the zoom meeting discussions.

Definitely, Vasco and Sasha (Gimelbrant) are role models to follow in the scientific world. They are always able to find a solution to every problem, share their scientific knowledge and raise important questions, which end up leading to important results.

I would also like to single-out Clara F. Alves-Pereira, who never gave up on this work and introduced me to the R programming language. Despite the long-distance, her support was important for an excellent working environment and to finish what seemed like a never-ending project.

To all lab members for sharing their time with me and their precious help for this project.

To Elsa Seixas, who has been my support by being always present, available, and truly believing in me.

And finally, a very special thank goes to my family, friends, and all close friends, for their support and encouragement. Your help was very important for the success of this work.

This project has taken me a lot of effort, and I would have faced many more difficulties without the help of the people mentioned above.

Table of contents

List of figures

Figure 1. 4.Hematopoiesis models. The classical model assumes that HSCs are a homogeneous population of cells. All blood cells come from the HSC pool through a differentiation process (lineage commitment) that is characterized by discrete intermediate progenitors, each with reduced self-renewal ability. The HSC sits at the top of the hierarchy, and the binary branching represents the cell fate decisions during lineage commitment direction. The first step of lineage commitment is the separation of MPPs into CMPs and CLPs. CLPs give rise to lymphocytes, whereas CMPs differentiate into MEPs and GMPs. MEPs are progenitors of megakaryocytes/platelets and red blood cells. GMPs produce granulocytes, macrophages, and dendritic cells. With the improvement of HSC isolation, new cell surface markers, and a large collection of works based on single-cell or limiting dilution cell transplantation, new findings on HSC were revealed. This led to a revised version of the classical model. This model includes a new branching decision, the first lineage separation that produces CMPs and LMPPs. CMPs give rise to MEPs and GMPs. LMPPs produce CLPs and also GMPs. Additionally, a direct shortcut into the megakaryocytic lineage was suggested (dashed lines). As these two models cannot explain the heterogeneity of the HSC compartment, a new model was proposed. In this model, it is assumed that HSC is a heterogeneous pool of cells and this heterogeneous behavior of HSC is an intrinsic feature epigenetically established early in development. Hematopoiesis is defined as a continuous flow of differentiation and emergence of lineage trajectories independent of each other without obvious hierarchical boundaries. The classical Waddington landscape is used to visualize this model. HSC, hematopoietic stem cell; MPP, multipotent progenitor; CMP, common myeloid progenitor; CLP, common lymphoid progenitor; LMPP, lymphoid-primed multipotent progenitor; MEP, megakaryocyte-erythroid progenitor; GMP, granulocyte-macrophage progenitor; MgK, megakaryocytes; RBC, red blood cells; Mac, macrophage; Gr, granulocyte; DC, dendritic cell; NK, natural killer cell. 35

Figure 4. 20. B clones expanded *in vitro* show more genes with clonal-specific AI than B cells differentiated from a single HSC *in vivo*. (A) Representative dot plots of pairwise comparison of AI between different Abelson-immortalized B-cell clones. Red circles signal the genes for which differential AI remained statistically significant after QCC correction on the binomial test. The total number of these genes per comparison is shown above each plot. The Pearson's coefficient correlation for all AI pairwise comparisons is also shown at each dot plot's upper left corner. Mean abundance levels (mean TMM-normalized counts) are continuous greyscale colors. (B) Correlogram with pairwise comparisons of Abelson-immortalized B-cell clones. Pearson's coefficient correlation of AI for all pairwise comparisons between samples. Pearson's coefficient is represented in the upper right corner within each square, and the number of genes with a significant differential AI in each pairwise comparison after applying QCC correction on the binomial test is also shown. (C) Two dot plots showing standard deviations (SD) of AIs for four monoclonal (x-axis) against four polyclonal (y-axis) HSC-derived B cell samples (left plot), and SD of AI for all four Abelson clones (x-axis) against the SD of AI for four polyclonal HSC-derived B cell samples (y-axis) (right plot). Whole-exome sequencing data were used to exclude transcripts with possible LOH. Dashed vertical and horizontal lines set arbitrarily at an AI SD of 0.15 represent the threshold above which genes were considered potentially intrinsically imbalanced. Mean abundance levels (mean TMM-normalized counts) are represented as binned greyscale colors. .. 95

Figure 4. 21. The strategy for identification of X chromosome inactivation escapees. (A) and (B) Allelic imbalance of X-linked genes for B and T cells, respectively. As a convention, *Xi allelic imbalance=1* means that the gene is 100% expressed from the inactive X-linked allele; *Xi allelic imbalance=0* means that only the active X-linked allele was detected. Dots represent genes with expression higher than 10 TMM-normalized counts, and only genes that were statistically different from the sample-corrected threshold at least once are shown. Yellow dots represent monoclonal samples; violet stroke-surrounded yellow dots denote statistical significance for that sample. Red dots represent the median of the allelic imbalance observed for polyclonal and control samples (otherwise excluded from this top panel to compare Xi allelic imbalance of

List of tables

List of abbreviations

AI - allelic imbalance
ASAR15 - asynchronous replication and autosomal RNA on chromosome 15
ASAR6 - asynchronous replication and autosomal RNA on chromosome 6
ASL - astrocyte-like cell
B6 - C57BL/6
CAST - Cast/Ei
CLP - common lymphoid progenitor
CMP - common myeloid progenitor
DC - dendritic cell
EMed - Escola de Medicina
ESC - embryonic stem cell
FACS - fluorescence-activated cell sorting
FCM - Faculdade de Ciências Médicas
FCT - Fundação para a Ciência e a Tecnologia
FM - Faculdade de Medicina
FPKM - fragments per kilobase of exon model per million mapped reads
FSE - Fundo Social Europeu
Gfap - glial fibrillary acidic protein
GMP - granulocyte-macrophage progenitor
Gr - granulocyte
HSC - hematopoietic stem cell
HSPC - hematopoietic stem and progenitor cell
IGC - Instituto Gulbenkian de Ciência
IL - interleukin
iPSC - induced pluripotent stem cell
IT - intermediate-term
Lin - lineage population
LINE-1 - long interspersed nuclear element-1
LMPP - lymphoid-primed multipotent progenitor
lncRNA - long non-coding RNA
LOH - loss of heterozygosity
LSD1 - lysine-specific demethylase-1
LSK - $Lin^{-}Sca\text{-}1^{+}cKit^{+}$
LT - long-term
m - maternal
Mac - macrophage
MDR - multi-drug-resistance
MEP - megakaryocyte-erythroid progenitor
MgK - megakaryocytes

MPP - multipotent progenitor
NGS - next-generation sequencing
NK - natural killer cell
NMS - Nova Medical School
NPC - neural progenitor cell
NSC - neural stem cell
OR - olfactory receptor
ORBEA - Órgão Responsável pelo Bem-Estar dos Animais
ORF - open reading frame
p - paternal
PCA - principal component analysis
PhDOC - Inter-University Doctoral Programme in Ageing and Chronic Diseases
Pkp3 - plakophilin 3
POCH - Programa Operacional Capital Humano
QCC - quality control correction
RBC - red blood cell
RMAE - random monoallelic autosomal expression
RPKM - reads per kilobase of exon model per million reads
SLAM - signalling lymphocyte activation molecule
SLE - systemic lupus erythematosus
SNP - single nucleotide polymorphism
ST - short-term
TLR7 - Toll-like receptor 7
TMM - trimmed mean of M
TPM - transcripts per million
t-SNE - t-distributed stochastic neighbor embedding
UC - Universidade de Coimbra
UM - Universidade do Minho
UNL - Universidade NOVA de Lisboa
WES - whole-exome sequencing
Xce - X-controlling element
XCI - X chromosome inactivation
Xm - X chromosome of maternal origin
Xp - X chromosome of paternal origin

Resumo

Nos organismos eucarióticos diplóides, a maioria dos genes são expressos bialelicamente. No entanto, existem excepções em que, ao nível das células, a expressão ocorre num padrão monoalélico que resulta de uma transcrição diferencial dos alelos de base epigenética. Existem três classes de expressão monoalélica regulada por mecanismos epigenéticos: *imprinting* de origem parental, inactivação do cromossoma X (XCI*) e expressão aleatória monoalélica autossómica (RMAE). Populações enviesadas obtidas a partir de ensaios de transplante de uma única célula revelaram que o conjunto de células estaminais hematopoiéticas (HSCs) é heterogéneo, reflectindo as diferenças epigenéticas de células individuais. Segundo um modelo em que os padrões da expressão específica de alelos são estabelecidos durante a diferenciação de células estaminais embrionárias e são propagados depois de forma estável através de divisões celulares, as HSCs carregam genes (e alelos) com marcas epigenéticas estáveis. A análise a nível clonal dos estados epigenéticos das células estaminais é necessária para entender a sua heterogeneidade e diversidade. Nesta tese, avaliamos pela primeira vez a persistência de estados epigenéticos entre os alelos no sistema hematopoiético *in vivo* usando o desequilíbrio da expressão alélica como ferramenta de leitura. O trabalho baseou-se na criação de um sistema hematopoiético monoclonal em ratinho por transplante de uma única HSC e no subsequente estudo da progenia linfóide emergente por análise transcriptómica de todo o genoma.

Nas células hematopoiéticas resultantes de uma única HSC, verificámos que a XCI é mantida de forma estável após extensa proliferação e diferenciação, enquanto a vasta maioria dos genes autossómicos não estão sob RMAE. Assim, os paralelismos recorrentes na literatura entre XCI e RMAE são enganosos, porque estes dois fenómenos não têm a mesma estabilidade e serão regulados por diferentes mecanismos. Além disso, demonstramos que esta abordagem clonal com base num sistema sem manipulação genética pode ser uma estratégia para estudar a XCI específica de tecidos *in vivo*. Por fim, um padrão de RMAE foi encontrado num número raro de genes (14 genes, <0,2% do total) em células linfóides resultantes de uma única HSC, indicando que esses padrões já estavam presentes na HSC original usada no transplante. No entanto, o número de genes com RMAE em células que passaram por etapas de diferenciação é

muito menor do que o número relatado anteriormente em estudos usando linhagens celulares clonais *in vitro* sem diferenciação extensa (~2–15%). Para conciliar estas observações, propomos que a maioria dos padrões de RMAE são meta-estáveis, isto é, passíveis de eliminação e restauração em diferentes estados de diferenciação.

* As abreviaturas na língua inglesa foram mantidas.

Abstract

In diploid eukaryotic organisms, most genes are expressed biallelically. However, there are exceptions where the expression occurs in a monoallelic pattern that results from a differential allele-specific transcription based on the different epigenetic marking of the two alleles. At the level of cells, there are three classes of monoallelic expression regulated by epigenetic mechanisms: parent-of-origin imprinting, X chromosome inactivation (XCI), and random autosomal monoallelic expression (RMAE). Biased repopulations obtained from single-cell transplantation assays revealed that the pool of hematopoietic stem cells (HSCs) is heterogeneous, reflecting the epigenetic differences of individual cells. According to a model in which the allele-specific expression patterns are established during differentiation in embryonic stem cells and are stably propagated through cell divisions, it is assumed that HSCs carry genes (and alleles) with these stable epigenetic marks. Therefore, the analysis of epigenetic states in the stem cell population at the clonal level is necessary to understand its heterogeneity and diversity. Here we evaluated for the first time the persistence of allele-specific epigenetic states in the hematopoietic system *in vivo* using allelic imbalance as a readout. We created a monoclonal hematopoietic system in mice by single HSC transplantation and then analyzed the emerging lymphoid progeny using a genome-wide transcriptomics approach.

We revealed that in the single-HSC derived hematopoietic cells, XCI is stably maintained through extensive proliferation and differentiation, whereas the vast majority of autosomal genes lack the stable clonal patterns of random monoallelic expression. This finding shows that the recurrent parallels between XCI and RMAE are misleading, suggesting that different mechanisms underlie these two classes of monoallelic expression. Additionally, we show that this *in vivo* clonal approach, which is free of genetic manipulation, can replace the artificial strategies that have been used to study tissue-specific XCI. Finally, stable allele-specific expression patterns were found in a rare number of genes (14 genes, <0.2%) in the progeny of a single HSC, indicating that these patterns were already present in the original HSC used for transplantation. However, the number of genes with stable monoallelic expression in cells that underwent differentiation steps is much lower than the numbers previously reported in studies

using clonal cell lines *in vitro* without extensive differentiation (~2–15%). To reconcile these observations, we propose that most allele-specific expression patterns in autosomal genes are metastable and can be erased and reestablished at different differentiation stages.

1. Introduction

1.1. Monoallelic expression

A diploid eukaryotic organism inherits one allele from each parent. In most cases, the two alleles of each autosomal gene are expressed simultaneously and at similar levels in cells. This pattern is called biallelic expression (Chess, 2016; Eckersley-Maslin and Spector, 2014; Gendrel et al., 2016). However, there are several exceptions to this rule in which one allele is transcriptionally downregulated or even silent and the other allele is upregulated or fully transcribed, leading to allelic imbalance (AI) or even monoallelic expression (**Figure 1.1**). This AI can be due to genetic polymorphisms in *cis*-regulatory elements that influence the efficiency at which a gene is transcribed and genetic modifications such as copy number variation of portions of the genome. Here we will focus on a different class of AI, which is not due to genetic differences in the sequences of the alleles but epigenetic mechanisms (Chess, 2016).

The epigenetics-based AI comprises three classes of monoallelic expression. In **Parent-of-origin imprinting** (**Figure 1.1**), monoallelic expression results from epigenetic imprints deposited during gametogenesis in the male and female germlines, leading to constitutive monoallelic expression from the same allele in the entire organism and more rarely in specific tissues (Reik and Walter, 2001). This monoallelic expression was discovered in 1984, when nuclear transplantation was used to generate embryos with two sets of chromosomes either from two maternal pronuclei or two paternal pronuclei. These embryos (and control embryos) were transplanted to pseudo-pregnant mice and it was shown that the presence of both maternal and paternal genomes is essential for normal development (Barton et al., 1984; McGrath and Solter, 1984; Surani et al., 1984). Approximately 100–180 imprinted genes were found in mice and 50–100 in humans (Babak et al., 2015; Baran et al., 2015; Barlow, 1995; Chess, 2016). Imprinting has relevance for development and diseases (Ferguson-Smith and Bourc'his, 2018). The abnormal inheritance of imprinted genes is associated with indistinguishable deletions in a cluster of imprinted genes present in the human chromosome 15, leading to the Prader-Willi syndrome (a developmental disease with defects in growth control and brain function), if the perturbation is in the paternal allele, or Angelman's syndrome (a disease associated with severe mental retardation), if the perturbation is in the maternal

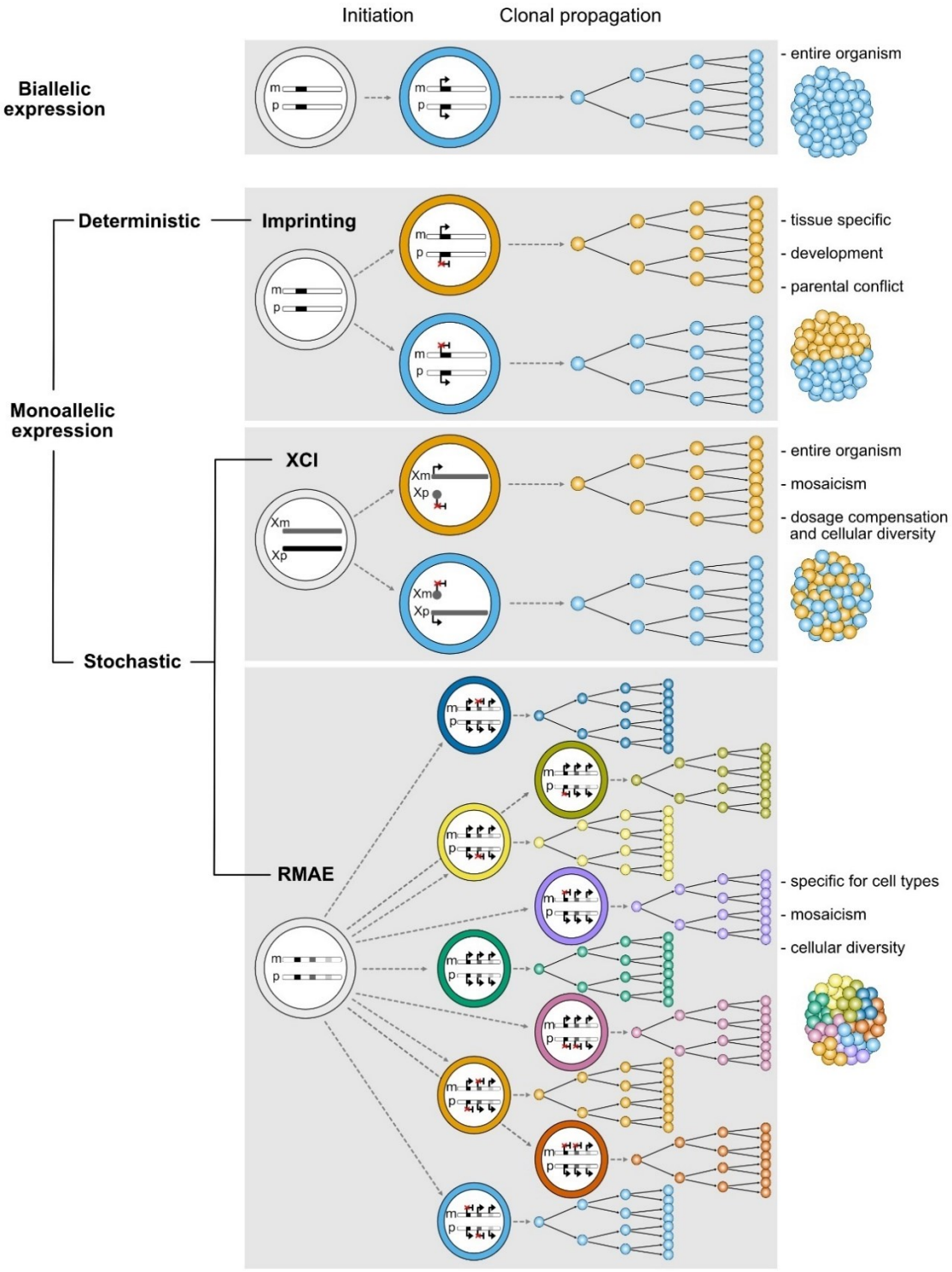


Figure 1. 1. Schematic representation illustrating biallelic expression features and different types of random monoallelic expression. Monoallelic expression can be divided into cases based on a deterministic choice of allelic expression, including imprinting genes, and cases based on stochastic choice, including X-chromosome inactivation (XCI) and random monoallelic autosomal expression (RMAE). Xp, X chromosome of paternal origin; Xm, X chromosome of maternal origin; p, paternal autosome; m, maternal autosome. Adapted from our review (Barreto et al., 2021).

allele. For each impacted gene, the loss of one allele is not compensated by the expression of the other allele (Nicholls et al., 1998). Despite the risk of genetic diseases, as all mutations in the single active allele of imprinted gene will be dominant, the persistence of this phenomenon for over 125 million years of mammalian evolution suggests that imprinting must provide advantages (Renfree et al., 2009). The evolution of imprinting remains an open question and several competing theories for the origin of imprinting have been proposed, namely the two most prevalent ones: parental conflict and coadaptation. The first hypothesis affirms that there will be selection for a paternal allele that enhances transfer of nutrients to the fetus, regardless of the fitness of the mother (Haig and Westoby, 1989). The second hypothesis states that the selection of the maternal allele expression is favored when it increases the adaptative offspring and maternal traits, resulting in higher offspring fitness (Wolf and Hager, 2006). These two hypotheses are not mutually exclusive (Renfree et al., 2009).

The other two broad classes are characterized by random monoallelic expression choice followed by stable mitotic transmission, so that different cells from the same organisms express mostly or exclusively the paternal or maternal alleles. These classes are **X chromosome inactivation** (XCI) (Lyon, 1961) and **random monoallelic autosomal expression** (RMAE) (Gimelbrant et al., 2007) (**Figure 1.1**). XCI, also named "Lyonisation," was first described in 1961 by mouse geneticist Mary Lyon while studying coat-color variegation in mice. Lyon concluded that one of the two X chromosomes must be randomly silenced in each female cell, leading to monoallelic expression of the remaining active X chromosome genes. She hypothesized that XCI allows the XX female to keep the same expression level ratio of the X-linked versus autosomal genes found in the XY male. XCI occurs early in development, around the time of implantation, resulting in a mosaic expression of the paternal and maternal X chromosomes in female tissues. Failure to establish this random monoallelic expression is also associated with genetic disorders affecting the brain, bone, blood, ears, heart, liver, kidney, retina, skin, and teeth. Generally, in the presence of recessive mutations, males will be more affected than women and manifest more severe phenotypes, because males carry only one X chromosome and the presence of the mutant allele will lead to the manifestation of the

phenotype. In contrast, women heterozygous for the mutation in a X-linked gene will have a mosaic pattern of expression due to XCI, ensuring that half of the cells will express a normal allele (the other half will express the mutated allele), which prevents or attenuates the manifestation of the deleterious phenotype. It should be stressed that heterozygous females would probably not be as exposed to the mutated allele as hemizygous males even in the absence of XCI, since they carry two X chromosomes and the recessive mutant allele would be compensated by the wildtype allele (Migeon, 2020). However, the escape from XCI, which is observed for some X-linked genes, increases the protein dosage (compared to XY cells) due to its expression from both X chromosomes. The extra dosage of proteins linked to immunity can be advantageous for females if it provides more protection against infectious diseases, but it also can make females more vulnerable to autoimmunity (Migeon, 2020; Mousavi et al., 2020). For example, it is known that the dosage of Toll-like receptor 7 (TLR7) is a key factor in systemic lupus erythematosus (SLE), an autoimmune disorder with a female bias. Additionally, a single-cell study of *TLR7* revealed that it escapes XCI in the B cell compartment, which leads to higher protein expression, IgG antibody production, and responsiveness to TLR7 ligands (Souyris et al., 2018).

RMAE, the third class of monoallelic expression, shares some similar features with the X-linked genes under XCI. The choice of allelic expression in these two classes of monoallelic expression is stochastic, giving rise to mosaicism (with some cells expressing one allele and other cells expressing the opposite allele in the somatic tissues). On the other hand, genome-wide studies showed that RMAE occurs at the gene level across all chromosomes without preferential location or clustering, like genomic imprinting, rather than in the chromosome-specific manner that characterizes XCI. It is assumed that the choice of allelic activity is set during development and is stably kept through cell division. This expression can lead to differences in gene expression levels and, in the context of heterozygosity, can contribute to cell identity and cellular heterogeneity (also known as phenotypic diversity at the cell level) (Gendrel et al., 2014; Gimelbrant et al., 2007; Zwemer et al., 2012). Unlike parent-of-origin imprinting, which does not create

cellular diversity at the organismal level, allele-specific expression resulting from XCI or RMAE is observed at the single-cell or clonal cell line levels (**Figure 1.1**).

1.1.1. Random Monoallelic Autosomal Expression

In the 1960s, the first cases of autosomal genes under monoallelic expression were discovered, namely the antigen receptor genes (B cell (immunoglobulins) and T cell receptors (Pernis et al., 1965)). This phenomenon would subsequently be described as "allelic exclusion", a term that became associated with the antigen receptor genes, which undergo a unique process of DNA rearrangement mediated by the RAG recombinases called V(D)J recombination (Hozumi and Tonegawa, 1976). The B and T cell receptor genes were for several decades the only known cases of RMAE. However, in 1994, a different gene family, unrelated to the immune system, the 1,296 murine odorant receptors expressed by sensory neurons, was found to be under RMAE (Chess et al., 1994; Zhang and Firestein, 2002). Furthermore, in the 2000s, RMAE was extended to the gene family of protocadherins, which are expressed in the central nervous system (Esumi et al., 2005; Tasic et al., 2002). This RMAE of large gene families of the nervous and the immune system generates individual cell identity and intercellular diversity. In addition to these gene families, other studies reported several individual autosomal genes with monoallelic expression affecting a wide range of functions and cell types. These studies are summarized in **Table 1.1**.

Until 2007, RMAE was considered to apply only to a few genes encoding proteins involved in the immune and nervous systems. That view was changed by data from the first genome-wide identification of allele-specific transcription of genes, which was performed in clonal human B-lymphoblastoid cell lines. Using an SNP (single nucleotide polymorphism)-sensitive microarray (by hybridizing cDNAs to SNP arrays), ~5–10% of the assessed genes were found to be under RMAE. These genes encode proteins with a wide range of diverse functions but include a large fraction of genes encoding cell surface proteins. Reduced gene expression was associated with monoallelic expression. Genes subjected to monoallelic expression were more likely to be near presumed regulatory conserved sequences related to accelerated evolution. A conservative extrapolation of these data suggested that more than 1,000 human genes are subject to

Table 1. 1. List of autosomal genes under random monoallelic expression described in studies focused on single genes. The table shows the year of the study, the studied tissue or cell type, and the type of work (*in vitro* or *in vivo*). Adapted from our review (Barreto et al., 2021).

Gene	Cell type / tissue	Species	*In vitro*/ *in vivo*	Year	References
immunoglobulin receptor genes	B and T lymphocytes	rabbit mouse	*in vivo*	1965 1976 1985	(Cebra and Goldstein, 1965; Goverman et al., 1985; Hozumi and Tonegawa, 1976; Pernis et al., 1965)
olfactory receptors (*OR*) genes	sensory neurons	mouse	*in vivo*	1994	(Chess et al., 1994)
HUMARA (human androgen receptor)	colonic crypts	human	*in vivo*	1995	(Endo et al., 1995)
Ly49 receptor genes	natural killer cells	mouse	*in vivo*	1995	(Held et al., 1995)
interleukin genes (*IL2*, *IL4*, *IL5*, *IL10*, *IL13*)	T cells	mouse	*in vitro*	1998, 2000, 2006	(Bix and Locksley, 1998; Calado et al., 2006; Holländer et al., 1998; Kelly and Locksley, 2000)
Pax5	early progenitors and mature B cells	mouse	*in vitro*	1999	(Nutt et al., 1999a)
VRi2	sensory neurons of the vomeronasal system	mouse	*in vivo*	1999	(Rodriguez et al., 1999)
Nubp2, Igfals, and *Jsap1*	bone marrow stromal cells and hepatocytes	mouse	*in vitro*	2001	(Sano et al., 2001)
Variable lymphocyte receptors (VLRs) genes	lymphocytes	lamprey	*in vivo*	2004	(Pamcer et al., 2004)
Protocadherin genes	Purkinje cells	mouse human	*in vitro* / *in vivo*	2002 2005 2006	(Esumi et al., 2005; Kaneko et al., 2006; Tasic et al., 2002; Wang et al., 2002)
Tlr4	B cells	mouse	*in vitro*	2003	(Pereira et al., 2003)
KIR genes	natural killer cells	human	*in vitro*	2003	(Chan et al., 2003)
Cd4	CD4+ lymphocytes	mouse	*in vitro*	2004	(Capparelli et al., 2004)
p120 catenin	pre-B clonal cell lines	mouse	*in vitro*	2005	(Gimelbrant et al., 2005)
	lymphoblastoid lines	human			

Gfap (*glial fibrillary acidic protein*)	cortical astrocytes	mouse	*in vitro*	2008	(Takizawa et al., 2008)
rDNA loci	lymphoblasts	human	*in vitro*	2009	(Schlesinger et al., 2009)
Krt12	limbal stem cells	mouse	*in vivo*	2010	(Hayashi et al., 2010)
IGF2BP1	B cells	human	*in vitro*	2011	(Thomas et al., 2011)
ASAR6	P175 cell line (derived from HTD114 fibrosarcoma cell line)	human	*in vitro*	2011	(Stoffregen et al., 2011)
Cubilin	renal proximal tubules and small intestine	mouse	*in vivo*	2013	(Aseem et al., 2013)
ASAR15	P268 cell line (derived from HTD114 fibrosarcoma cell line)	human	*in vitro*	2015	(Donley et al., 2015)
Gata3	hematopoietic stem cells and early T-cell progenitors	mouse	*in vitro* / *in vivo*	2015	(Ku et al., 2015)
FOXP2	B lymphoblastoid cell lines and clonal T-cell lines	human	*in vitro* / *in vivo*	2015	(Adegbola et al., 2015)
Bcl11b	T cells	mouse	*in vitro* / *in vivo*	2018	(Ng et al., 2018)

random monoallelic expression (Gimelbrant et al., 2007). With the emergence of new technologies of transcriptome genome-wide analysis, other studies would confirm that the frequency of autosomal genes with random monoallelic expression is higher than what was previously believed (**Table 1.2**). In 2012, to test whether RMAE is also widespread in the mouse genome, like in humans, murine clonal B-lymphoblastoid cell lines derived from a hybrid F1 line were genome-wide analyzed by the SNP-sensitive microarray. 15.6% of autosomal genes were revealed to be under RMAE. It was demonstrated that genes under RMAE in the mouse display a wide distribution across the genome and diverse functions. Many of these genes are orthologous for human organisms, suggesting conservation for genes under RMAE between species. RMAE was

Table 1. 2. A summary of reports based on genome-wide transcriptomics analysis in different cell types. The table describes the type of experimental assay, studied species and genotype (if applicable), number of analyzed clones, and the observed percentage of genes under RMAE. Adapted from our review (Barreto et al., 2021).

Cell type	Experimental assay	Species	Genotypes	% of RMAE	#analyzed clones	References
Lymphoblastoid cells (*in vitro*)	SNP-sensitive microarrays	human	NA	5-10	12	(Gimelbrant et al., 2007)
		mouse	129S X CAST; Balb/c X C57BL/6J	15.6	11	(Zwemer et al., 2012)
Fibroblasts (*in vitro*)	SNP-sensitive microarrays	mouse	129S X CAST	2.1	2	(Zwemer et al., 2012)
	RNA-seq	mouse	CAST X 129S	0.52-1.9	6	(Pinter et al., 2015)
Neural stem cells (*in vitro*)	SNP-sensitive microarrays	human	NA	1.4-2.0	9	(Jeffries et al., 2012)
	RNA-seq	mouse	C57BL/6 X JF1	2.4	4	(Li et al., 2012)
	SNP-sensitive microarrays	human	NA	0.63	3	(Jeffries et al., 2016)
	RNA-seq	mouse	C57BL/6 X JF1	4.6	4	(Branciamore et al., 2018)
Neural progenitor cells from embryonic stem cells (*in vitro*)	RNA-seq	mouse	C57BL/6 X CAST	3.0	6	(Eckersley-Maslin et al., 2014)
			129S X CAST	2.5	8	(Gendrel et al., 2014)
Embryonic stem cells (*in vitro*)	RNA-seq	mouse	C57BL/6 X CAST	0.5	6	(Eckersley-Maslin et al., 2014)
iPSC (*in vitro*)	SNP-sensitive microarrays	human	NA	0.88	2	(Jeffries et al., 2016)
Neural stem cells from iPSC (*in vitro*)	SNP-sensitive microarrays	human	NA	0.65-0.84	2	(Jeffries et al., 2016)
Astrocyte-like cells (*in vitro*)	RNA-seq	mouse	C57BL/6 X JF1	6.4	4	(Branciamore et al., 2018)

also shown for 2.1% of genes in murine fibroblast clones (Zwemer et al., 2012). Another study also confirmed monoallelic expression, revealing 0.52–1.9% of genes with this expression pattern in mouse tail-tip fibroblasts. RNA-sequencing (RNA-seq) was performed from primary and clonal cell populations obtained from hybrid F1 mouse reciprocal crosses. This strategy enabled the identification of allelic imbalance due to genotype, imprinting, or RMAE. It was observed that genes under RMAE presented lower expression levels but longer transcripts and gene bodies than balanced genes, and enrichment for cell surface and adhesion terms (Pinter et al., 2015).

Several studies were also performed using neural cell lines. In 2012, genome-wide allelic expression assessment using RNA-seq and hybrid F1 mouse line revealed that around 2.4% of autosomal genes in clonal neural stem cells (NSCs) are subjected to RMAE with a portion of genes belonging to the glutathione superfamily. Additionally, it was observed that genes with monoallelic expression are 30–35% less expressed than those with biallelic expression. They maintain their monoallelic expression pattern when NSCs are differentiated into neurons and astrocytes. Some of these genes are potentially relevant for diseases (Li et al., 2012). Furthermore, in 2012 different clonal NSCs derived from the human cerebral cortex, striatum, and spinal cord were analyzed by SNP-sensitive microarrays and revealed 1.4–2.0% of genes under RMAE. The identified monoallelic genes belong to transmembrane glycoproteins, neurodevelopmental proteins, and transcription factor binding sites. After differentiation of these clones into neurons and glia, the allelic expression status was maintained. The data of Gimelbrant et al. (2007) was also confirmed, showing that monoallelic genes have reduced gene expression and are more likely to be located close to accelerated evolutionary non-coding sequences (Jeffries et al., 2012). The same group, in 2016, using the same strategy, investigated the time point in the development when the stochastic allelic choice is made. Allelic expression imbalances were profiled in human neural progenitor cells (NPCs) before (NSC line clones) and after (iPSC (induced pluripotent stem cells) epigenetic reprogramming and after neuralization of iPSCs back to a NSC state (rosette-like neural stem cells). It was found that, in NSCs, the 0.63% of genes that are under RMAE generally switch to the biallelic state after reprogramming to iPSCs. In iPSCs,

0.88% of genes showed RMAE, and at least 34% of these genes were biallelically expressed before reprogramming. It was concluded that these genes are potentially *de novo* monoallelically expressed genes that emerged at the pluripotent state. After neuralization, 0.65% of expressed genes were characterized by RMAE, and many of them were not monoallelically expressed in the original neural stem cell line (Jeffries et al., 2016). Another study performed by RNA-seq of clonal lines of neural progenitor cells derived from F1 hybrid mouse embryonic stem cells (ESCs) revealed 2.5% of genes under RMAE. Most of these genes showed low expression levels, and when compared to biallelic clones, monoallelic clones showed lower abundance. The monoallelic genes in NPCs are known to be involved in cell adhesion and organ development. In contrast to the finding of Jeffries et al. (2016), in this study, it was observed that most monoallelically expressed genes in NPCs were biallelic or not expressed in original ESCs, suggesting that the monoallelic status is established during differentiation toward NPCs. Additionally, after 15 NPC clone passages, the monoallelic expression patterns of genes were maintained. The same observation was made when NPCs differentiated to astrocytes (Gendrel et al., 2014). Also, in 2014, another allele-specific RNA-seq screen was performed for RMAE during differentiation of F1 hybrid mouse ESCs to NPCs. A 5.6-fold increase in monoallelic expression during differentiation was noticed, from 0.49% in embryonic cells to 3.0% in progenitor cells, with only 2.0% of monoallelically expressed genes overlapped between stem and progenitor cell lines, indicating that monoallelic expression is acquired upon lineage commitment early in development, similarly to what was found by Gendrel et al. (2014). 8% of genes in NPC monoallelic clones showed the same expression level as in biallelic clones, indicative of transcriptional compensation, and 15.4% showed lower levels (Eckersley-Maslin et al., 2014). A more recent work, in which NSCs were differentiated *in vitro* into astrocyte-like cells and genome-wide transcriptome sequencing was performed, identified 4.6% of genes under RMAE in NSCs and 6.4% of in astrocyte-like cells. It was observed that 3.3% of genes with RMAE are common between two developmental stages, and 21.8% and 26.0% of genes with RMAE are specific for NSCs and astrocyte-like cells, respectively. Genes under RMAE in undifferentiated cells are enriched for genes encoding proteins with cell division and DNA replication functions. In contrast, genes with RMAE in

differentiated cells include genes responsible for the differentiated activity of neural cells (ion channels, transporters, and cell surface components). 44% of genes with RMAE in astrocyte-like cells were not expressed in undifferentiated cells (Branciamore et al., 2018). Several candidate genes under RMAE previously found to be under monoallelic expression were analyzed in 200 murine NPCs before and after differentiation from ESCs, and it was found that allele-specific expression is established during differentiation; however, only two genes in undifferentiated stem cells were analyzed (Marion-Poll et al., 2021). It was also confirmed that monoallelic expression for genes previously identified as RMAE (Gendrel et al., 2014) is stably maintained once stochastically established in five NPC clones that differentiated to astrocytes. Additionally, by analyzing two monoallelic candidates in all clones, it was observed that monoallelic expression is not always associated with low expression. In some cases, RMAE is associated with down-regulated protein levels and could have functional consequences (Marion-Poll et al., 2021). After comparing different studies, genes under RMAE within the same tissue have considerable overlap suggesting that different studies are consistent (**Figure 1.2**).

RMAE is highly tissue-specific (Gendrel et al., 2016; Marion-Poll et al., 2021), an observation supporting the notion that monoallelic expression patterns are established during differentiation (Branciamore et al., 2018; Eckersley-Maslin et al., 2014; Gendrel et al., 2014). In studies that compared different cell types and different species such as humans and mice, it was observed that, although in different cell types different genes under RMAE were expressed, these genes were involved in the same type of activity (Nag et al., 2013); and genes under RMAE are highly conserved between human and mouse (Nag et al., 2015). The absence of overlap observed between different cell types could be explained by cell-type-specific gene expression, meaning that a gene, which is monoallelically expressed in one cell type, could even be not expressed in another one.

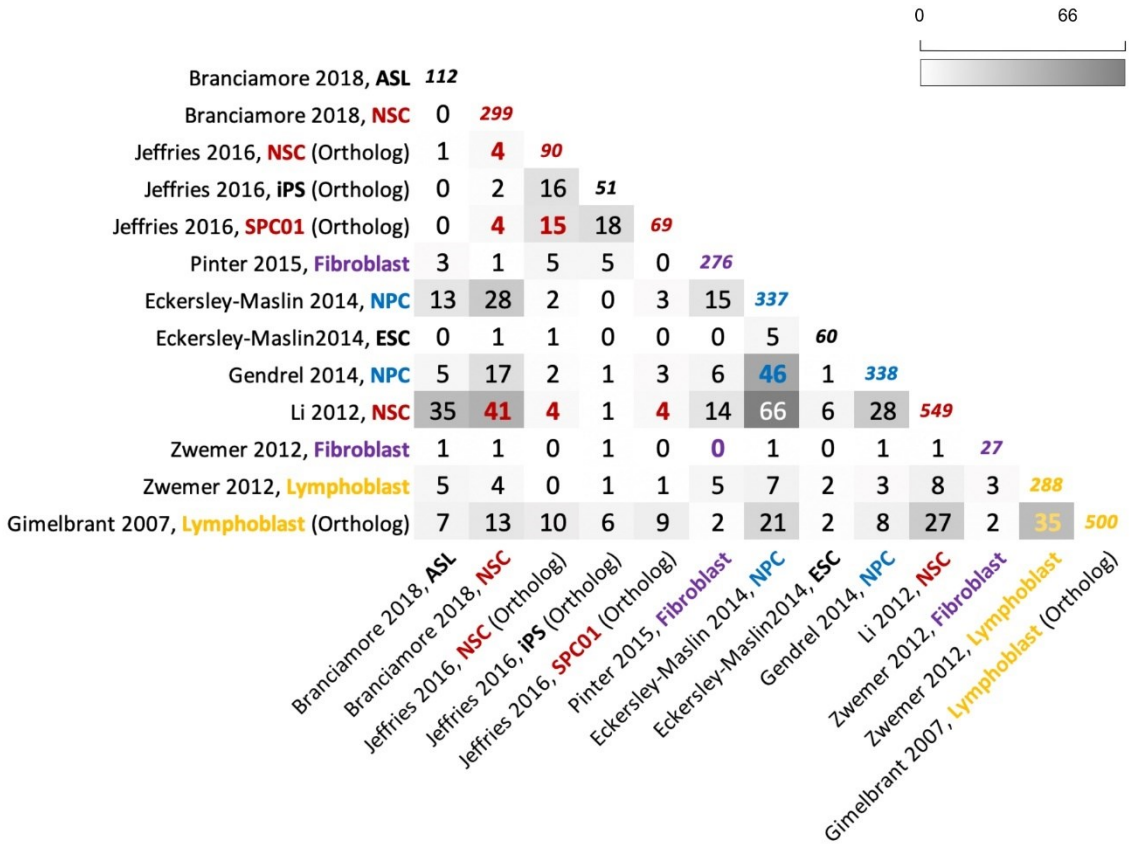


Figure 1. 2. Half-matrix showing all pairwise intersections of autosomal gene collections identified as random monoallelically expressed in the genome-wide studies described in Table 1.2. (except Jeffries et al., 2012, which is not publicly available). ASL, astrocyte-like cells; NSC, neural stem cells; NPC, neural progenitor cells; ESC, embryonic stem cells; SPC01, clonal neural stem cells (before epigenetic reprogramming); iPSC, induced pluripotent stem cells after epigenetic reprogramming of SPC01. Note that "NPC" in Jeffries et al., 2016 is derived from iPSCs. Colors represent instances where a different cell/tissue type was studied more than once. To obtain intersections, gene ids were briefly manually curated for inconsistencies (e.g., gene name-to-date conversions when the originally provided data were in Microsoft Excel format). All gene sets were then parsed with the gprofiler2 R package (Raudvere et al., 2019) for gene id consistency, using transcript ids as query whenever possible and ENSEMBL gene ids as target (performed July 12th, 2021). Orthology conversion (from human to mouse) was performed with the same package for datasets involving human data. For Gimelbrant et al. (2007) and Zwemer et al. (2012) gene collections, MAE classes I, II, and III were used to retrieve genes with RMAE, and for Gendrel et al. (2014), the "NPC_random_catalog" classification was retrieved as RMAE. Adapted from our review (Barreto et al., 2021).

1.1.2. Clonal versus dynamic RMAE

Deng and colleagues performed single-cell RNA-seq on 269 individual cells dissociated from *in vivo* F1 (CAST/EiJ x C57BL/6J) mouse preimplantation embryos (stages from oocyte to blastocyte). Abundant (12–24%) RMAE in the mammalian embryonic cells was reported, which was also confirmed in individual adult mouse liver cells and cultured mouse fibroblasts. However, by pooling cells from the same embryo, and considering that each embryo is a clone of cells, almost all monoallelic expression observed at the level of individual cells was removed. It was concluded that this abundant monoallelic expression is highly variable among newly divided cells from the same embryo and is not due to fixed monoallelic expression propagated through cell divisions. Instead, it is rather independent, dynamic, and consistent with transcriptional bursting models (Deng et al., 2014). Two other works using the same strategy on the human reference lymphoblastoid cell line GM12878 and human primary fibroblast cell line reported that most genes in single cells are monoallelically expressed from one or the other allele and a few genes display biallelic expression at a given time point. This dynamic allele expression, which changes from cell to cell, as if each corresponds to an independent snapshot, could be explained by transcriptional bursting. It was also observed that allele-specific expression is correlated with the abundance of the transcript, i.e., a large portion of highly expressed genes were transcribed from both alleles, whereas genes expressed at low levels were monoallelically transcribed (Borel et al., 2015; Marinov et al., 2014).

Most of the studies mentioned above were performed on single cells that are not clonally related (Borel et al., 2015; Marinov et al., 2014). In this approach, the calculation of clonal and dynamic fractions in monoallelic expression is impossible. Reinius and colleagues used single-cell RNA-seq on clonal primary F1 (CAST x B6) murine fibroblasts and freshly isolated $CD8^+$ from a male human donor vaccinated with a yellow fever vaccine to explore the contribution of clonal and dynamic monoallelic expression. The percentage of genes under RMAE in primary fibroblasts and CD8 T cells was 13% and 60–85%, respectively. However, 95% of this expression was dynamic, with fewer than 1% of total genes showing clonal monoallelic expression, and most of these rare genes

had low expression levels (Reinius et al., 2016). In 2021, the same group proposed again that transcriptional bursting (the number of bursts in time units) can explain the monoallelic expression observed in single-cell RNA-seq data. (Larsson et al., 2021).

The dynamic RMAE shows varied allelic expression patterns among cells from the same clone or the same embryo that is due to independent stochastic transcriptional bursting, meaning that at any specific time, RNA from only one allele is often transcribed and is detectable (Deng et al., 2014; Larsson et al., 2021; Reinius et al., 2016). The major difference between clonal and dynamic RMAE is the long-term and short-term persistence of monoallelic expression in the cell, respectively. Clonal RMAE is stably maintained and mitotically inherited once established (Gimelbrant et al., 2007; Zwemer et al., 2012). Dynamic RMAE is not conserved in daughter cells after divisions and is ephemeral, even during the lifespan of a cell, resulting from the intrinsic stochasticity of transcription (Deng et al., 2014; Larsson et al., 2021; Reinius et al., 2016).

Although single-cell RNA-seq can be used to study cell heterogeneity, cell states, rare cell types, cells in early development, and cancer cells, the major handicap of this technique is the difficulty to distinguish biological variation from technical noise (Chen et al., 2019; Kim et al., 2015; Marinov et al., 2014). Compared to bulk RNA-seq, single-cell RNA-seq uses low amounts of starting material, limiting the efficiency and uniformity of RNA transcription into cDNA and consequently restricting its representation in the library and capture, producing more noise and higher technical variations. In addition, single-cell RNA-seq data contain many missing values (the gene is expressed but not detected by RNA-seq (Kharchenko et al., 2014)) and dropouts (the gene is detected at an intermediate or high level in one cell but not detected in another cell (Kharchenko et al., 2014)), introducing many false positives in allelic expression profiles (Marinov et al., 2014). Because of these limitations, appropriate sequencing protocols, efficient quality control, and sophisticated and complex analytical tools are necessary to eliminate this noise. A major part of allele-specific expression defined as stochastic can be explained by technical noise since the used approaches have difficulty correctly distinguishing the true biological allelic bias expression from the technical variability present in RNA-seq data (Kim et al., 2015). Furthermore, single-cell RNA-seq

presents complex issues in sequencing protocols and data analysis. This technique is less suitable than bulk RNA-seq in allele-specific expression studies. A major feature of genes under RMAE is low expression levels (Eckersley-Maslin et al., 2014; Gendrel et al., 2014; Gimelbrant et al., 2007; Jeffries et al., 2012; Li et al., 2012; Reinius et al., 2016) and a low number of RNA molecules can introduce amplification bias. The current single-cell RNA-seq technology can detect extremely low RNA levels, on average <1 mRNA per cell at the population level. Taken these considerations together, it is currently unclear if dynamic RMAE is a true phenomenon or an artifact (Marinov et al., 2014). Additionally, multiple possible cell states, cell cycle or mRNA processing, and splicing can exist within a cell population, leading to cell-to-cell heterogeneity of gene expression (Chen et al., 2019; Marinov et al., 2014).

1.1.3. Mechanisms

Not much is known about the RMAE regulating mechanisms. Parallels between RMAE and XCI were drawn to interpret better how RMAE is regulated (initiated, propagated, and maintained) since XCI is more well-defined and these two types of monoallelic expressions share some features, namely the randomness and clonal propagation through cell divisions (Chess, 2016; Gendrel et al., 2016; Goldmit and Bergman, 2004). However, there are many differences between these two types of monoallelic expression (Barreto et al., 2021). In contrast to XCI, in which only maternal or paternal chromosome is expressed in each cell, genes under RMAE in independent clonal populations originated from the same progenitor cell can be either monoallelically expressed from maternal/paternal allele or biallelically expressed, or even completely silent. It is known that XCI is established early in development, whereas for RMAE it is unclear when this pattern of monoallelic expression is established in development; it can occur early or late, when the gene starts to be expressed. In the case of olfactory receptor genes, the stochastic choice of monoallelic and monogenic expression of a unique allele from the pool of olfactory receptor (OR) gene family occurs in the maturing olfactory sensory neurons of the mouse olfactory epithelium (Magklara et al., 2011). For XCI, the initiation of inactivation starts with the stochastic transcription of the key

regulator non-coding RNA *Xist* from the chromosome, which will be inactivated. For RMAE, it is unclear how the maternal or paternal allele starts being stochastically expressed. In some cases, both alleles are in the default silent state before the expression of one of the alleles (ex. olfactory receptor genes, immunoglobulin genes, and *Ly49* genes). In other cases, both alleles are initially expressed and then one allele is silenced (ex. *KIR* genes) (Anderson, 2014; Degl'Innocenti and D'Errico, 2017; Vettermann and Schlissel, 2010). Which epigenetic regulation is used to mitotically propagate and keep stable monoallelic expression through several cell expansion and differentiation is also not clear. The immunoglobulin and T cell receptor genes, which were the first cases of RMAE to be discovered and have been intensively investigated, are exceptional cases of monoallelic expression because of the process of V(D)J recombination that occurs in developing lymphocytes. The mitotic stability of the patterns of RMAE that antigen receptor genes display is partly due to somatic genetic alterations, a scenario that does not apply to all other genes under RMAE. Similarly, in olfactory sensory neurons, once the monoallelic expression of olfactory receptor genes is established, it is stably maintained. However, these cells are terminally differentiated, which is not the case of the majority of examples of RMAE (Degl'Innocenti and D'Errico, 2017).

An enormous amount of work has been performed to understand the master regulation of RMAE (see below). These studies were based on clonal cell lines propagated *in vitro* and phenotypically stable as they do not undergo substantial differentiation programs. However, monoallelic expression is observable only in clonal cell populations because in non-clonal cells this expression is masked at the cell population level, i.e., it appears biallelic. For most tissues, in part because tracking monoclonal cell populations *in vivo* is challenging, the regulatory mechanism underlying remains unclear.

Negative feedback

The cases of allelic exclusion of the immunoglobulins, T cell receptors, and OR genes depend on a negative feedback: the expression of one allele leads to the inhibition of

recombination (antigen receptor genes) or expression (OR genes) of the other allele. For immunoglobulins, once the first allele is successfully rearranged by V(D)J recombination and the protein expressed by this allele is present at the cell surface, the rearrangement of the second allele is inhibited. However, if the rearrangement of the first allele was not successful, the second allele has a chance to recombine (Alt et al., 1984; Kitamura and Rajewsky, 1992). Each cell has a time window to rearrange the alleles. If the second allele is also unsuccessfully rearranged, the cell undergoes apoptosis (Chi et al., 2020; Vettermann and Schlissel, 2010). This feedback inhibition of V(D)J recombination ensures that each mature B cell expresses only one monospecific antigen receptor in the diverse repertoire of antigens. However, supposing the feedback mechanism is slow, after the rearrangement of the first allele the second allele has time to rearrange within a time window, producing cells expressing both alleles. In the same way, if the time window is limited, the second allele would not have time to recombine, and fewer cells with biallelic expression would be formed (Barreto et al., 2021). T cell receptors have a similar mechanism of negative feedback during V(D)J recombination (Aifantis et al., 1997) (**Figure 1.3**). Similarly, the feedback mechanism of olfactory receptor genes leads to the monoallelic expression of a single allele of a single gene in each neuron from the large repertoire of the olfactory receptor gene family, which is present in clusters over several chromosomes, resulting in a functional olfactory system (Serizawa et al., 2003). In the olfactory epithelium, all OR alleles are initially repressed by constitutive heterochromatin. During neuronal differentiation, a single OR allele is selected in a stochastic manner for demethylation by lysine-specific demethylase-1 (LSD1), an enzyme that catalyzes demethylation of H3K9, (temporarily expressed at the time window of OR gene choice) and released from this transcriptional inhibition. After that, a feedback mechanism is activated that results in downregulation of LSD1 to avoid the activation of additional OR genes and maintain the stable expression of the chosen OR gene, promoting the maturation of olfactory neurons. If a non-functional OR gene was chosen, LSD1 retains its activity, and a new OR gene is selected for expression until a functional receptor is eventually chosen. (Degl'Innocenti and D'Errico, 2017; Nagai et al., 2016) (**Figure 1.3**). A negative feedback mechanism has not been described for the other genes under random monoallelic expression.

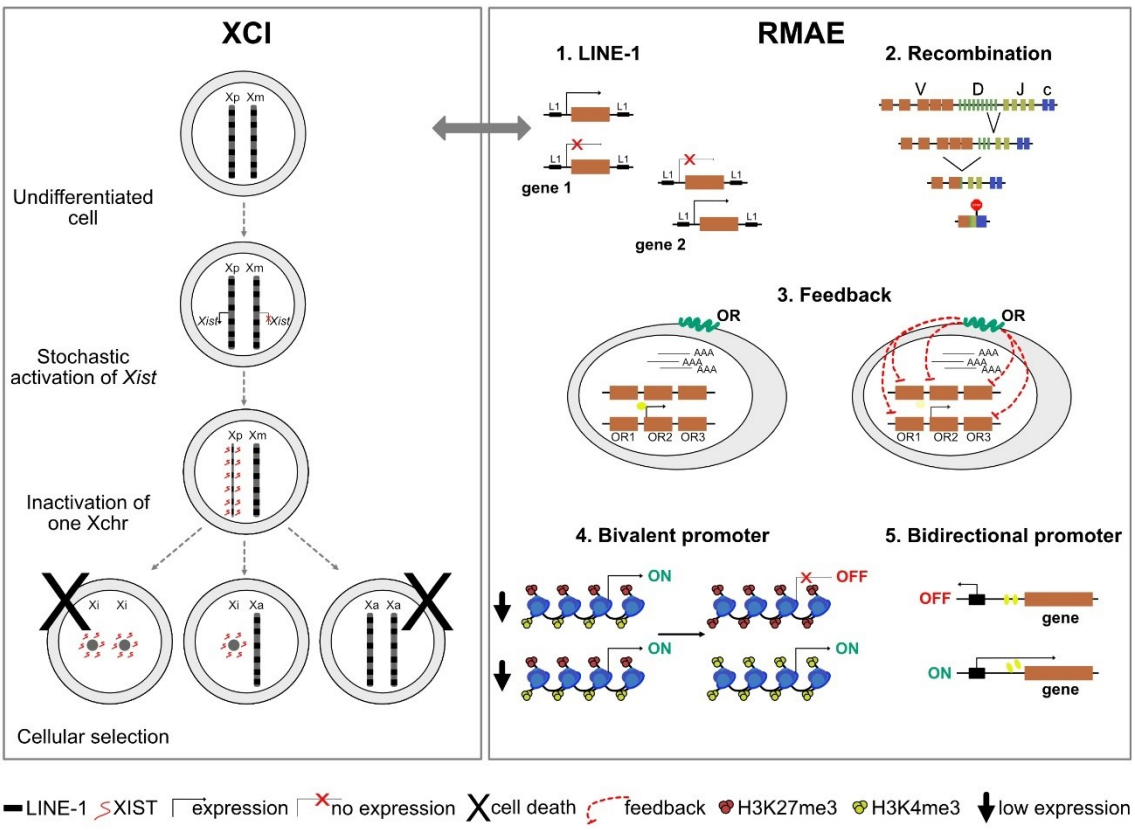


Figure 1. 3. A schematic representation of mechanisms responsible for X-chromosome inactivation (XCI) and mechanisms possibly responsible for RMAE. The gray arrow represents a potential parallel between XCI and RMAE associated with LINE-1 elements (L1). Adapted from our review (Barreto et al., 2021).

Nuclear organization

The nuclear organization is relevant for XCI. The inactivated X chromosome is usually found at the nuclear periphery or within the perinucleolar region. These regions are associated with heterochromatin, which can help the silencing of the inactive X chromosome by exposing the chromosome to heterochromatin factors and limiting the access to transcription factors (Chow and Heard, 2010). The nuclear position is also implicated in immunoglobulin and OR gene RMAE. The nuclear organization of immunoglobulin is altered during B-cell differentiation. Both alleles are closely associated with the nuclear periphery in very early B progenitor cells and not with

heterochromatin. Moreover, the productive allele is repositioned at a more central part during B-cell differentiation, while the non-productive allele occupies the nuclear periphery (Kosak et al., 2002; Skok et al., 2001). Mature sensory olfactory neurons have a unique inverted nuclear architecture. In these cells, olfactory receptor genes from different chromosomes are concentrated in a small number of compact *foci* (large domains of constitutive heterochromatin aggregates) around a large constitutive heterochromatin block, whereas the active OR allele is localized outside of olfactory receptor *foci* close to the more plastic facultative heterochromatin (Armelin-Corre et al., 2014; Clowney et al., 2012).

The nuclear organization was investigated in other examples of monoallelically expressed genes. For example, the astrocyte-specific marker *Gfap* (glial fibrillary acidic protein), monoallelically expressed in cortical astrocytes, showed differential nuclear positioning for the active and inactive allele with the active allele positioned more internally (Takizawa et al., 2008). However, six monoallelically expressed genes in neural progenitor cells were analyzed, and the expression status did not correlate with differential nuclear positioning of active and inactive genes (Eckersley-Maslin et al., 2014). Thus, aside from immunoglobulins, olfactory receptor genes, and the isolated example of *Gfap*, currently nuclear positioning does not seem to play a role in genes under RMAE, but this potential mechanism needs systematic investigation.

Bidirectional promoter switches

The class I MHC receptors expressed by natural killer cells are encoded by the *Ly49* genes in mice and the *KIR* genes in humans, and both are monoallelically expressed. This stochastic monoallelic expression is achieved using *cis* probabilistic bidirectional promoter switches that produce sense and antisense transcripts, which in turn activate or silence gene expression, respectively. However, the gene silencing or activation mechanism is different for *Ly49* and *KIR* genes. In the case of murine natural killer cells, the stochastic switch occurs at the distal bidirectional promoter. Thus, the *Ly49* genes are inactive, until sense non-coding transcripts produced by the distal promoter (Pro1)

activate the downstream promoter (Pro2); the antisense non-coding transcripts keep the state inactive (Saleh et al., 2002, 2004). In contrast, in *KIR* genes the stochastic switch occurs at the proximal promoter in human cells. Thus, the *KIR* genes are initially in an active state. Antisense transcripts from the proximal bidirectional promoter switch produce a piRNA that promotes an inactive state. When sense transcripts are produced, they only preserve the active state of the allele (Davies et al., 2007). Thus, the murine *Ly49* genes use sense transcripts to activate a state that is "off" by default, and *KIR* genes use antisense transcripts to inactivate a state that is “on" by default. No more examples of this type of regulation mechanism have been documented; it is unclear how frequent bidirectional promoters are associated with genes under RMAE (**Figure 1.3**).

Asynchronous replication

During asynchronous DNA replication, the active allele with open chromatin replicates early in the S phase, whereas the inactive allele with compact heterochromatin replicates in the late S phase. This is one of the features of XCI (Takagi, 1974). Asynchronous replication has also been associated with RMAE. For instance, it was shown that olfactory receptors genes in neurons undergo random asynchronous replication (Chess et al., 1994). Other examples of genes under RMAE were found to be asynchronously transcribed: *Il-2* in mouse T cells (Holländer et al., 1998), *p120* in mouse and human (Gimelbrant et al., 2005); and two human non-coding RNA genes, *ASAR6* (asynchronous replication and autosomal RNA on chromosome 6) (Donley et al., 2013; Stoffregen et al., 2011) and *ASAR15* (asynchronous replication and autosomal RNA on chromosome 15) (Donley et al., 2015). These two former large intergenic non-coding RNA genes will be discussed below in more detail. Chromosome-wide coordination of asynchronous replication was observed in mice (Singh et al., 2003) and human autosomal chromosomes (Ensminger and Chess, 2004), but the choice of the allele to be expressed in RMAE is not coordinated at the chromosome level. Different genes under RMAE on the same chromosome can be expressed independently from the maternal, paternal, or both alleles (Gimelbrant et al., 2007). Similarly, no clear correlation between monoallelically expressed genes and asynchronous replication was found in neural

progenitor cells (Gendrel et al., 2014). Antigen receptor *loci* were also associated with this distinctive replication of XCI. It was found that similarly to XCI, the pattern of asynchronous replication is lost in the morula, re-established at the time of implantation, and then clonally maintained. It was proposed that this predetermined replication feature is introduced on the allele, which is the early replicating one and selected for V(D)J rearrangement, and then retained as the cell propagates and undergoes differentiation (Mostoslavsky et al., 2001). This parallel with XCI was broken in subsequent work performed by the same group. This study proposed that the epigenetic mark for asynchronous replication on the kappa light chain is established not early in development but about the time of lymphoid commitment (Farago et al., 2012). Another study demonstrated that allelic exclusion of immunoglobulin receptor *loci* is not predetermined until V(D)J recombination and thus stands apart from XCI. Additionally, this work showed that the immunoglobulin heavy chain alleles rearrange independently (Alves-Pereira et al., 2014). In this way, asynchronous replication does not seem to be the characteristic of RMAE or an useful feature for predicting genes under monoallelic expression.

DNA methylation

Another epigenetic mark associated with allele-specific expression in XCI is DNA methylation. After *Xist* has been upregulated on the X chromosome that will be inactivated, repression of the second *Xist* allele (from active X chromosome) is maintained through DNA methylation of its promoter (Sado et al., 2004). This epigenetic mark was extensively explored in genes under RMAE. A strong association between DNA methylation and monoallelically expressed *loci* in NSCs was observed. Increased DNA methylation levels at the CpG sites located in the proximal promoter regions of the transcript start sites were found when monoallelically expressed genes were compared with the biallelically expressed ones (Jeffries et al., 2012). The same association between RMAE and increased methylation at the gene promoter was observed in iPSC-derived NSC clones. Furthermore, decreased DNA methylation at the gene body in monoallelically expressed genes was detected (Jeffries et al., 2016). In NPCs, it was

observed that while some monoallelic genes displayed higher levels of DNA methylation at their promoters than biallelic genes, others did not show a correlation between methylation levels and monoallelic expression. When cells were treated with 5-azacytidine (which inhibits DNA methyltransferases), leading to DNA demethylation, no significant reactivation of the inactive allele was observed (Gendrel et al., 2014). Similar results were also obtained in the same cell line and with treatment with 5-azacytidine (Eckersley-Maslin et al., 2014). Additionally, it was observed that DNA methylation is correlated with expression levels rather than directly with the allele-specific expression pattern, and that alleles which are initially unmethylated in ESCs become methylated only when the allele is silenced or show low expression after differentiation to NPCs. As CpG islands showed hypomethylation in clones with biallelic expression and ESCs, intermediate methylation was found in clones with monoallelic expression, and hypermethylation in clones where a gene was silenced or showed low expression levels. In the same work, it was shown that the silent allele of the monoallelically expressed *Bag3* gene was derepressed using decitabine (DNA methyltransferase inhibitor), suggesting that DNA methylation maintains the monoallelic expression of this gene (Marion-Poll et al., 2021). In another recent work, a screening-by-sequencing approach for reactivation of silenced alleles across 23 *loci* of mouse genome was performed. In this study, 43 drugs were tested in four monoclonal lines of pro-B cells. The small molecules studied are known to be involved in introducing or eliminating methylation and acetylation marks on histones and DNA. It was reported that 5-azacytidine and 5-aza-2'-deoxycytidine could reactivate the silenced allele in many *loci*, but this was not, however, dependent on the level of DNA methylation. It was suggested that DNA methylation plays a key role as a fine-tuning mechanism of RMAE, which regulates allele-specific transcription as a rheostat regulation rather than an on-and-off switch control. And for some *loci,* other mechanisms in addition to or instead of DNA methylation are responsible for RMAE maintenance (Gupta et al., 2021). In these two more recent works, especially in the study of screen, it was revealed that DNA methylation is one of the key mechanisms of RMAE maintenance. However, until now, it is not known how and when this mark is established.

Histone modifications

Chromatin signature is another epigenetic signal associated with maintaining monoallelic expression after its initiation. Specific histone modifications are correlated with active (H3K4me3 and H3K36me) and silent (H3K9me3 and H3K27me3) chromatin states. Genes subject to XCI retain silent chromatin marks associated with heterochromatin (H3K9me3, H4K20me3, and H3K27me3). On the other hand, genes escaping XCI are enriched in active histone marks (H3K4me3, H3K9ac, and H3K9me1) (Goto and Kimura, 2009). In clonal NSCs, this correlation was also observed; monoallelically expressed genes showed enrichment in repressive marks (H3K27me3), whereas biallelically expressed genes showed enrichment in active marks (H3K4me3 and H3K9ac) (Jeffries et al., 2012). Similarly, a correlation between chromatin marks and monoallelic expression was also shown in NPCs. Biallelic clones showed an increase for active (H3K4me2/3 and H3K36me3) and a decrease for inactive marks (H3K9me3 and H3K27me3) when compared with monoallelic clones (Eckersley-Maslin et al., 2014; Gendrel et al., 2014). In recent work, the treatment of cells with histone deacetylase inhibitors (dacionostat and CUDC-101) increased the expression of the silent allele of the monoallelically expressed *Bag3* gene (Marion-Poll et al., 2021). Additionally, the presence of dual chromatin marks for active (H3K36me3) and inactive transcription (H3K27me3) on the gene body was proposed to be a signature of genes under RMAE (discussed in more detail later) (Nag et al., 2013). Despite different histone modifications being sufficient to distinguish active and inactive alleles, it is unclear if these chromatin marks are responsible for maintaining monoallelic expression through mitotic divisions and how these marks are introduced.

Long interspersed nuclear element-1

The long interspersed nuclear element-1 (LINE-1) transposon family, a DNA sequence frequent in X chromosomes, is associated with X chromosome inactivation. It was proposed that LINE-1 acts as booster elements promoting the spread of *Xist*, a key player in *cis*-acting in XCI, and gene silencing the X chromosome to be inactive (Lyon, 1998). It

is now known that *Xist* long non-coding RNA (lncRNA) does not interact directly with LINE-1 but rather exploits the three-dimensional conformation of the X chromosome. By high-resolution mapping, an anti-correlation between *Xist* enrichment and LINE-1 sequences was found. *Xist*, transcribed from the X-inactivation center, contacts first with sites proximal by their spatial conformation to the *Xist* transcription locus. After that, it follows a two-step mechanism, initially spreading to gene-rich regions depleted in LINE-1 sequences and then targeting gene-rich regions, which are known to be enriched for LINE-1 (Engreitz et al., 2013; Leung and Panning, 2014; Simon et al., 2013). Although it is unlikely that LINE-1 acts as way stations for *Xist*, these sequences could participate in the XCI differently. An RNA-FISH study revealed two classes of LINE-1, which participate in X chromosome silencing at different levels. Old LINE-1 (truncated) sequences are inactive before XCI, and they facilitate efficient gene silencing by assembling heterochromatic nuclear compartments induced by *Xist* into which the genes to be silenced become recruited. Young LINE-1 elements are expressed from both X chromosomes before XCI. After differentiation and XCI establishment, these sequences are expressed from the inactive X chromosome and silenced on the active X chromosome. These active elements participate in the spreading of silencing in regions that would otherwise be prone to escape (Chow et al., 2010) (**Figure 1.3**). Additionally, a strong positive association between local susceptibility to XCI and LINE-1 density was observed in studies using X-autosome translocations and *Xist*-transgene inducible systems (Loda et al., 2017; Tannan et al., 2014).

The involvement of LINE-1 elements in XCI raises the possibility that genes under RMAE may also show high levels of repetitive sequences. It was found that monoallelic autosomal genes are flanked by significantly higher densities of LINE-1 sequences, fewer CpG islands, and fewer SINE elements in their flanking regions than genes expressed biallelically. These LINE-1 sequences are less truncated and evolutionary more recent than those found in biallelically expressed genes (Allen et al., 2003). The olfactory receptor genes of the main nose and the vomeronasal type-1 and type-2 receptor genes of the vomeronasal organ are also embedded in LINE-dense regions (Kambere and Lane, 2009). In NSCs, a minimal increased length of LINE-1 was observed in monoallelically

expressed genes. But this difference was not significant compared to biallelically expressed genes. However, a significant decrease in the amounts of SINE repeats and an increase in amounts of long terminal repeats across the length of the transcript were observed (Jeffries et al., 2012). Additionally, in NPCs, monoallelic genes tend to be enriched in regions increased in LINE-1 and reduced in SINE (Gendrel et al., 2014).

Interestingly, the human autosomal genes *ASAR6* and *ASAR15* share many features with *Xist*. For example, these genes express large non-coding RNAs, display random monoallelic expression and replicate asynchronously in coordination with other linked monoallelic genes. Like *Xist*, *ASAR6* and *ASAR15* are transcribed from the later replicating allele. Deletion of either gene results in a late replication phenotype and structural instability of the respective chromosomes in *cis*. Additionally, disruption of either gene leads to the activation of the previously silent allele of the nearby monoallelic genes. Another similarity with *Xist* is the delayed replication timing of chromosomes upon ectopic integration of cloned genomic DNA containing *ASAR6* or *ASAR15*. All three large non-coding RNAs contain a high density of LINE-1 in the transcribed sequence. Also, *ASAR15* forms a chromosome-sized cloud, similarly to *Xist*, but in contrast, *ASAR6* does not coat the chromosome entirely (Donley et al., 2013, 2015; Stoffregen et al., 2011). Other differences exist between the X chromosome and autosomal non-coding RNAs. More recently, it was shown that the LINE-1 element present in *ASAR6* in antisense orientation controls the replicating timing of chromosomes (Platt et al., 2018). However, it is not known how frequently this type of gene silencing occurs in autosomes, and, in contrast to *Xist*, *ASAR6* is not expressed in all adult tissues (Stoffregen et al., 2011) and *ASAR15* in some cells shows biallelic expression (Platt et al., 2018).

LINE-1 has also been observed to be related with the nuclear organization. It is known that monoallelic expression of olfactory receptor genes is associated with a unique inverted nuclear architecture (silenced genes are localized close to the nucleic center in a constitutive heterochromatin block, in contrast to usual localization in the periphery of silenced genes) and these genes are enriched in LINE-1 sequences. A recent study reported that LINE-1 is transcribed in the olfactory epithelium, and LINE-1

retrotransposons establish aggregates around the central constitutive heterochromatin blocks and partially colocalize with the facultative heterochromatin only in olfactory neurons. It was suggested that LINE-1 retrotransposons participate in organizing the specific nuclear architecture of olfactory neurons (Ormundo et al., 2020). This observation also draws parallels with XCI since LINE-1 is associated with the organization of heterochromatic nuclear compartments in the X chromosome to be inactivated (Chow et al., 2010). The LINE-1 enrichment is probably the strongest parallel between XCI and random monoallelic expression (**Figure 1.3**).

Bivalent domains

Repressive (H3K27me3) and active (H3K4me3) chromatin marks were for years considered to be mutually exclusive. However, a novel chromatin modification pattern where repressive and active marks are simultaneously present in embryonic stem cells was discovered (Azuara et al., 2006; Bernstein et al., 2006). These bivalent domains are more enriched in ESCs than in differentiated cells. And they are associated with transcription factor genes expressed at low levels with functions in embryonic development and lineage differentiation. These bivalent domains tend to acquire either an active or a repressive mark during embryonic stem cell differentiation. It was proposed that bivalent domains silence developmental genes in ESCs, which are then activated during differentiation (Bernstein et al., 2006).

Interestingly, H3K27me3 (silent mark) and H3K36me3 (active mark) present along the gene body but physically segregated in different alleles, were shown to account for most of the distinction between autosomal monoallelic and biallelic genes (**Figure 1.3**). This chromatin signature was found in up to 20% of the ubiquitously expressed genes and over 30% of tissue-specific genes, and was suggested to be a general feature of RMAE. Interestingly, more than 80% of genes under RMAE in differentiated cells were marked by bivalent promoters (with the silent and active marks physically linked to the same sequence) and in progenitor cells. These genes, drivers of differentiation, play a crucial role in determining cell fate during development. By drawing parallels with embryonic

stem cells, it was speculated that bivalent genes with poised chromatin, which are silent upon reaching a point of lineage commitment, can turn into an active or inactive state independently for the two alleles, resulting in monoallelic expression. In this case, one of the alleles can become stably silent (and enriched with gene body repressive mark), while the other becomes stably active (and enriched with gene body active mark). There is also some probability that both alleles become active, resulting in a biallelic state. After this resolution, the state is locked (Nag et al., 2013). This feature was proposed to be used as a proxy to predict genes under RMAE. A classifier based on these bivalent chromatin marks was developed and validated for different human and mouse cell types (Nag et al., 2013, 2015). Bivalent domains were also observed in promoters of allelically skewed genes enriched for development and cell differentiation players in mouse embryonic fibroblast lines (Savol et al., 2017).

1.1.4. Consequences

The obvious biological consequence of RMAE for the organism is phenotypic diversity at the cellular level (**Figure 1.1**). A textbook example of the functional importance of RMAE is the DNA rearrangement mechanism that generates a diverse repertoire of cells, each with a single and unique antigen receptor, which is necessary for proper functionating of the immune system because it avoids dual specificities (Chi et al., 2020). Another example of this type of functional importance is the RMAE of odorant receptors, which results in the production of different sensory olfactory neurons, each with only one expressed receptor (Chess et al., 1994), a feature necessary for odor detection and guiding the axons to the proper glomeruli (Wang et al., 1998). This role of RMAE in the formation of a precise topographic map was revealed by genetic experiments. When a deletion or non-sense mutation was introduced in the *Op2* olfactory receptor gene fused to *lacZ* in mouse (to trace the cells), the convergence of the axons into the glomeruli was altered. They wandered instead of acquiring a specific location. Additionally, a set of substitution experiments was performed where a given odorant receptor sequence was replaced by the sequence of another receptor, similarly fused to *lacZ*. Axons were

projected to locations different from the original odorant receptor glomeruli (Wang et al., 1998).

RMAE also contributes to phenotypic diversity for other genes. Even if such diversity for a given gene is not as complex as that associated to the antigen and olfactory receptor families, the potential for diversity of the combination of genes under RMAE is considerable. In the presence of heterozygosity, the biallelic expression will produce phenotypically identical cells, whereas RMAE will produce three types of cells: maternal monoallelic expression, paternal monoallelic expression, or biallelic expression. When we consider more than one heterozygous gene with monoallelic expression, the number of combinations of different expression patterns in a single cell will speedily grow in the order of 3^n (where n is the number of RMAE genes with polymorphic alleles) (**Figure 1.1**). Additionally, a group of human RMAE genes based on chromatin signature (Nag et al., 2013) showed a higher nucleotide diversity than biallelically expressed genes. This increased genetic diversity is mediated by increased mutation and recombination rates and balancing selection, indicating that RMAE has an advantage and evolved to boost phenotypic diversity at the cellular level (Savova et al., 2016). Many genes recently uncovered to be under RMAE are overrepresented in cell surface proteins, indicating their role in providing unique cellular identity (Gendrel et al., 2014; Gimelbrant et al., 2007; Pinter et al., 2015). Furthermore, monoallelic genes tend to be cell type-specific and their frequency increases upon cell differentiation, suggesting their significance in controlling regulatory pathways of lineage commitment and development (Branciamore et al., 2018; Eckersley-Maslin et al., 2014; Gendrel et al., 2014; Jeffries et al., 2012).

Monoallelically expressed genes usually have lower gene expression when compared to the biallelically expressed ones (Eckersley-Maslin et al., 2014), so even in the absence of heterozygosity, monoallelic expression can lead to cellular diversity resulting from differential gene expression and consequent differential protein dosage (Chess, 2013). Furthermore, this differential expression of genes expressed from one and two alleles has potential in fine-tune gene dosage, which can be important for cell development or response to stimulus (Gendrel et al., 2014; Marion-Poll et al., 2021).

Despite the advantages of cellular phenotypic diversity, random monoallelic expression can present disadvantages. In the presence of heterozygosity, monoallelic gene expression exposes an organism to the risks associated with the unmasking of recessive mutations, contributing to disease. Examples of such genes under monoallelic expression were found in several studies. In experiments performed by Gendrel et al. (2014) in NPCs, several genes involved in human autosomal-dominant disorders were found to be under RMAE, such as *Eya1* and *Six1*, *Eya4*, *Bag3*, *SNCA*, and *Cal9a3*, which are implicated in branchio-oto renal (BOR) syndrome and deafness (Kumar et al., 1998; Ruf et al., 2004), deafness syndrome (Wayne et al., 2001), childhood muscular dystrophy (Selcen et al., 2009), Parkinson's disease (Polymeropoulos et al., 1997), and multiple epiphyseal dysplasias (Bönnemann et al., 2000), respectively. The RMAE of these genes is mitotically stable in clonal neural progenitor cells. RNA-FISH was performed in the kidney and eye, and it was shown that *Eya1* and *Eye4* tend to be monoallelically expressed *in vivo* during the development of organs known to be affected in these disorders (Gendrel et al., 2014).

In NPCs, the *Dfna5* gene, implicated in deafness (Van Laer et al., 1998), and the *Bag3* gene, associated with muscular dystrophy (Selcen et al., 2009), were shown to be monoallelically expressed (Eckersley-Maslin et al., 2014). Another gene, *MYO6*, associated with deafness (Melchionda et al., 2001), is also monoallelically transcribed (Gimelbrant et al., 2007). In addition, genes associated with muscular dystrophy *Hsp8* (Ghaoui et al., 2016) and *TTN* (Gerull et al., 2002) were found to be under monoallelic expression (Gendrel et al., 2014; Jeffries et al., 2012). Additional examples of genes such as *FOXP2* (Adegbola et al., 2015) implicated in speech and language disorder (Lai et al., 2001), *OTX2* (Jeffries et al., 2012) in dystrophia of retinal pigment epithelium (Vincent et al., 2014), *Gja1* (Li et al., 2012) in oculodentodigital dysplasia (Paznekas et al., 2009) and *Gli2* (Zwemer et al., 2012) in hypopituitarism (Arnhold et al., 2015), *RBFOX1* (Jeffries et al., 2016) in autism (Sebat et al., 2007), *Agc1* (Wang et al., 2007) in arrested psychomotor development, hypotonia, and seizures (Wibom et al., 2009), *Tbx5* gene (Gui et al., 2017) in Holt-Oram syndrome (Basson et al., 1994, 1997) and *HoxB* genes

(Savol et al., 2017) in cancer (Li et al., 2019) were also revealed to be under random monoallelic expression.

Correct gene dosage of random monoallelic genes can have important implications. There are two genes under random monoallelic expression associated with neurodegenerative diseases. *APP,* which was identified as monoallelically expressed in the first genome-wide study of RMAE in B cells (Gimelbrant et al., 2007), is associated with Alzheimer's disease (Mullan et al., 1992). Moreover, *SNCA,* which was identified as monoallelically expressed in different studies in neural progenitor cells (Eckersley-Maslin et al., 2014; Gendrel et al., 2014) and B cells (Gimelbrant et al., 2007), is associated with Parkinson's disease. In Alzheimer's disease, excess of APP results in the formation of amyloid plaques (Rovelet-Lecrux et al., 2006). In contrast, in Parkinson's disease, the increase of SNCA leads to the formation of Lewy bodies (Singleton et al., 2003). It has not been explored if these genes are monoallelically expressed in relevant brain regions *in vivo*. However, if this is true, the dysregulation of monoallelic expression of these genes with aging could lead to the manifestation of the disease (Gendrel et al., 2016). Thus, random monoallelic expression can cause pathological conditions either through the dosage difference between one or two expressed alleles or by expressing one out of two functionally different alleles. However, without a comprehensive statistical analysis, it is important to stress that it is unclear if the association of RMAE to disease is real or spurious.

Jeffries et al. (2013) have found an overrepresentation between genes under RMAE previously identified in clonal NSCs (Jeffries et al., 2012) and candidate risk genes from association studies for schizophrenia, which were taken from the Genetic Association Database (Becker et al., 2004). Genetic association studies are used to correlate candidate genes with disease or genetic variation. Furthermore, a higher number than predicted of monoallelically expressed genes in this cell type was found at the copy number variation datasets associated with autism and schizophrenia (Jeffries et al., 2013).

A more recent study investigated 200 newly derived NPC clones from mouse ESCs, before and after differentiation, revealing the expression of 12 genes previously identified to be under RMAE (Gendrel et al., 2014; Gimelbrant et al., 2007) that are involved in development and associated with diseases. This set of genes includes *App*, which was previously found to be monoallelically expressed in NPC and B-lymphoblastoid clones, but in hippocampal mouse neurons showed biallelic expression when analysed *in vitro* by RT-PCR or *in vivo* by pyrosequencing. However, highly variable patterns and degrees of allelic imbalance were observed between clones (Marion-Poll et al., 2021).

The importance and implications of random monoallelic expression *in vivo* have not been addressed yet. Most works that studied the correlation of RMAE with disease show that monoallelic expression has a particular role in brain function and development, but probably because most experiments were performed in clonal neural cells, such as neural progenitor cells differentiated from embryonic stem cells (Eckersley-Maslin et al., 2014; Gendrel et al., 2014; Marion-Poll et al., 2021) and neural stem cells (Jeffries et al., 2012, 2016; Li et al., 2012; Wang et al., 2007).

1.2. Hematopoietic stem cells

Hematopoietic stem cells (HSCs) were first identified in 1961 by showing that the injection of bone marrow cells into lethally irradiated mice leads to the development of hematopoietic colonies in the spleens of recipient mice (Till and McCulloch, 1961). Since then, HSCs have been extensively explored. An enormous effort has been made to isolate pure HSCs and identify their function, key molecular pathways, and regulation mechanisms. Different approaches based on fluorescence-activated cell sorting (FACS) and characteristic cell surface markers, including signalling lymphocyte activation molecules (SLAM) and/or vital dye staining, have been developed to improve HSC enrichment and their progenitors (Adolfsson et al., 2001; Akashi et al., 2000; Goodell et al., 1996; Kiel et al., 2005; Kondo et al., 1997; Okada et al., 1992; Osawa et al., 1996).

HSCs inhabit the bone marrow. They have the unique self-renewal capacity to maintain the pool of stem cells. They can progressively give rise to all functional hematopoietic cells through extensive proliferation and differentiation (Cheng et al., 2020; Mayle et al., 2013). They sustain the hematopoietic system in the right cell type and cell number by producing all the blood cells, i.e., the lymphoid blood cells (B cells, T cells, natural killer (NK) cells, and dendritic cells) and myeloid blood cells, which encompass platelets, erythrocytes, basophils, neutrophils, eosinophils, macrophages, and monocyte-derived dendritic cells. In contrast to other multipotent progenitor cells, only HSC can self-renew for long enough to maintain the hematopoietic system for a lifetime (Cheng et al., 2020; Dick, 2003; Mayle et al., 2013; Reya, 2003; Schroeder, 2010). They are a very rare cell population; only 1 in 12,500–25,000 adult mouse bone marrow cells is an HSC (Kiel et al., 2005; Uchida et al., 2003; Yang et al., 2005).

Using cell isolation protocols that define the expression of cell surface markers and transplantation and colony assays, which characterize the functional and differentiation ability of the HSCs, a classic model of hematopoietic tree based on immunophenotyping was developed around the year 2000. This model was proposed to explain better the relationship of HSCs with progenitor cells and their differentiation steps (Akashi et al., 2000; Kiel et al., 2005; Kondo et al., 1997; Manz et al., 2002; Morrison et al., 1997; Yang et al., 2005). In this linear branching model (**Figure 1.4**), HSCs occupy the top of a hierarchy. Multipotent long-term HSCs (LT-HSCs) can originate the same daughter cell or differentiate into discrete multipotent, oligopotent, and subsequently unipotent progenitor cell stages in a stepwise manner by several subsequent binary branching decisions. It is assumed that LT-HSCs are a homogenous cell population. LT-HSCs have a long-term reconstitution capacity (>3 months) with the production of all blood cells with the initial differentiation into short-term HSCs (ST-HSCs). These cells have short-term reconstitution potential (<1 month) due to the reduced self-renewal capacity and subsequently differentiate into multipotent progenitors (MPPs). After this, the first bifurcation originates oligopotent common lymphoid progenitors (CLPs) and common myeloid progenitors (CMPs). CLPs have lymphoid potential and give rise to B, T, and NK cells. On the other hand, CPMs undergo a second bifurcation and originate granulocyte-

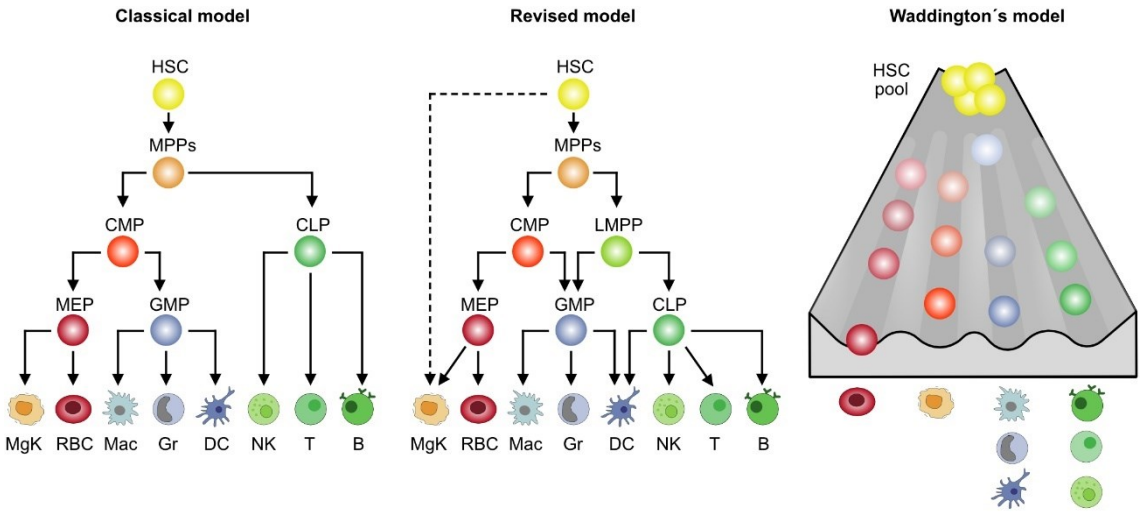


Figure 1. 4.Hematopoiesis models. The classical model assumes that HSCs are a homogeneous population of cells. All blood cells come from the HSC pool through a differentiation process (lineage commitment) that is characterized by discrete intermediate progenitors, each with reduced self-renewal ability. The HSC sits at the top of the hierarchy, and the binary branching represents the cell fate decisions during lineage commitment direction. The first step of lineage commitment is the separation of MPPs into CMPs and CLPs. CLPs give rise to lymphocytes, whereas CMPs differentiate into MEPs and GMPs. MEPs are progenitors of megakaryocytes/platelets and red blood cells. GMPs produce granulocytes, macrophages, and dendritic cells. With the improvement of HSC isolation, new cell surface markers, and a large collection of works based on single-cell or limiting dilution cell transplantation, new findings on HSC were revealed. This led to a revised version of the classical model. This model includes a new branching decision, the first lineage separation that produces CMPs and LMPPs. CMPs give rise to MEPs and GMPs. LMPPs produce CLPs and also GMPs. Additionally, a direct shortcut into the megakaryocytic lineage was suggested (dashed lines). As these two models cannot explain the heterogeneity of the HSC compartment, a new model was proposed. In this model, it is assumed that HSC is a heterogeneous pool of cells and this heterogeneous behavior of HSC is an intrinsic feature epigenetically established early in development. Hematopoiesis is defined as a continuous flow of differentiation and emergence of lineage trajectories independent of each other without obvious hierarchical boundaries. The classical Waddington landscape is used to visualize this model. HSC, hematopoietic stem cell; MPP, multipotent progenitor; CMP, common myeloid progenitor; CLP, common lymphoid progenitor; LMPP, lymphoid-primed multipotent progenitor; MEP, megakaryocyte-erythroid progenitor; GMP, granulocyte-macrophage progenitor; MgK, megakaryocytes; RBC, red blood cells; Mac, macrophage; Gr, granulocyte; DC, dendritic cell; NK, natural killer cell.

macrophage progenitors (GMPs) and megakaryocyte-erythrocyte progenitors (MEPs). GMPs produce granulocytes (basophils, neutrophils, and eosinophils) and monocytes (macrophages and monocyte-derived dendritic cells), and MEPs produce megakaryocytes that give rise to platelets, erythrocytes, or red blood cells.

1.2.1. Heterogeneity of HSCs

The classical model of hematopoiesis oversimplifies the complexity of hematopoietic stem and progenitor cells, and advances in technologies led to a new interpretation of hematopoiesis (Cheng et al., 2020; Schroeder, 2010). Studies on single HSCs found that these cells form a heterogeneous pool, as individual HSC can produce different ratios of mature blood cells, i.e., individual HSCs show distinct biases toward the myeloid or the lymphoid lineages differentiation (Benveniste et al., 2010; Dykstra et al., 2007; Muller-Sieburg et al., 2002, 2004; Sieburg et al., 2006).

Muller-Sieburg et al. followed the behavior of individual clonally derived HSCs using a long-term serial repopulation assay. Different HSC clones showed clearly distinguishable repopulation patterns. Although the reconstitutions were long-term, the extent and kinetics of hematopoietic repopulation were heterogeneous. Daughter HSCs from multiclonal grafts showed biased lineage contributions. However, daughter cells derived from individual clones had a homogeneous behavior, they were remarkably equivalent to each other in the extent and kinetics of repopulation, contributing similarly to the myeloid or lymphoid lineages. This homogenous pattern is inheritable, because it is still observed in secondary reconstitutions. It was suggested that, although the daughter HSCs within each clone behaved similarly, the pool of HSCs in adult bone marrow is heterogeneous. This heterogeneity is not generated continuously but early in development, and it is largely predetermined (Muller-Sieburg et al., 2002). Later, the same group studied the myeloid biased HSCs, which produce normal levels of myeloid precursors but reduced precursors for the T- and B- lymphocyte lineages. The lymphoid progeny of these cells expresses lower interleukin-7 (IL-7) receptor levels and fails to respond to IL-7 *in vitro*. This reduced response of the lymphoid progeny is epigenetically

programmed at the level of the HSCs and acts as a regulatory mechanism to decrease lymphopoiesis, thus affecting the number, composition, and function of the mature progeny (Muller-Sieburg et al., 2004). Further studies using systematic analysis of HSC heterogeneity assessed by clonal long-term repopulation patterns from transplantation assays of many individual HSCs confirmed that the HSC pool behaves in a way consistent with the idea of heterogeneous subpopulations of HSCs and that this heterogeneity is predetermined. It was found that the HSC compartment comprises a limited number of subpopulations of HSCs with predictable behaviors. This observation is inconsistent with a model in which HSCs are seen as a homogenous population of cells that produce a heterogeneous pattern by responding to different stimuli, because if so, each HSC clone should recreate the heterogeneity seen in the HSC compartment, and a continuous spectrum of clonal kinetic and self-renewal patterns would be observed. Together with previous data, it was concluded that the HSC pool comprises a limited number of different HSC types, each of them with predetermined and epigenetically fixed repopulation and self-renewal patterns (Sieburg et al., 2006).

Through a large-scale serial transplantation assay at the clonal level, Dykstra et al. found that the multipotent HSCs can be divided into four functionally different subpopulations (α, β, γ, and δ) with long-term reconstitutions but distinct self-renewal capacity and skewed ratios of myeloid and lymphoid generated populations. β cells display features conventionally associated with LT-HSCs, with long-term multipotentiality and self-renewal activity. γ and δ subpopulations show gradually less durable myeloid contributions and lack extensive self-renewal capacity, suggesting that these cells are developmentally downstream of α and β cells. α cells are myeloid biased subpopulations with self-renewal activity and can convert to β cells with lymphopoiesis capacity. β, γ, and δ cells can be accommodated into a classical model of hierarchical relationships. However, α cells are distinct and do not fit into this model. α and β subsets were associated with extensive self-renewal of the original cell transplanted and stable propagation of the original repopulating pattern to their progeny, with daughter cells from the same clone showing a high degree of similarity. This means that epigenetic

mechanisms in the cells intrinsically predetermine this HSC lineage bias before transplantation, possibly early in development (Dykstra et al., 2007).

Further insight about the molecular properties of HSC subpopulations was shown using the combination of the "side population" (made of cells with the capacity to exclude Hoechst dye by high multi-drug-resistance (MDR)-type transporters) and the CD150 marker. A very clear gradient of HSCs with distinct phenotypic, functional, and molecular characteristics was identified. This gradient describes populations of HSCs with high self-renewal capacity and myeloid bias or decreased self-renewal capacity and lymphoid bias phenotypes. Different responsiveness *in vivo* and *in vitro* to TGF-β1 was observed in myeloid-biased and lymphoid-biased subpopulations, enforcing their specific features and presenting a possible mechanism for the differential regulation of HSC subtype activation. The myeloid-biased HSCs exhibited the highest engraftment rate per mouse with single-cell reconstitutions and the highest overall contribution to peripheral blood generation. Results of secondary transplantation also demonstrated that myeloid-biased HSCs have a higher self-renewal capacity (Challen et al., 2010). Another group confirmed these results and identified different cell populations within the HSC pool based on CD150 expression. With extensive single-cell reconstitution assay, the cell population with high expression of CD150 was enriched in cells with long-term reconstitution capacity, great self-renewal potential, and multilineage potential. These cells have latent and hardly detectable myeloid contribution in primary recipients but progressive and multilineage reconstitution in secondary recipient mice. Other subsets of cells were identified with long-term reconstitutions but limited self-renewal potential and lymphoid bias. It was concluded that these differences in reconstitution activity are caused by intrinsic differences among HSCs (Morita et al., 2010).

Differences in self-renewal capacity with multilineage potential of long-term HSCs were also observed by the Iscove's group (Benveniste et al., 2003, 2010). Using the CD1149b marker and single-cell reconstitutions, intermediate-term HSCs (IT-HSCs) that self-renew longer than ST-HSCs but shorter than LT-HSCs were identified with erythroid and lymphoid lineage capacity. These IT-HSCs are numerically dominant over LT-HSCs and separable from LT and ST-HSCs. Because LT- and IT-HSCs are both quiescent, both

multipotent for erythro-lymphomyelopoiesis, and both present sufficient self-renewal capacity to produce detectable systemic grafts from a single cell, the key difference being the ability to sustain self-renewal in to the long-term, it was suggested that the first step in stem cell differentiation is not the loss of the mechanism that executes self-renewal but the loss of the mechanism that sustains self-renewal capacity in LT-HSCs. This heterogenous self-renewal capacity was attributed to inherent cellular characteristics rather than stochastic decisions made after transplantation (Benveniste et al., 2010).

More recently, heterogeneity in lineage output was confirmed by *in situ* using barcoding methods to clonally trace progenitors and stem cells, suggesting that this heterogeneity is not only associated with the emergency of hematopoiesis in transplantation assays but also characterizes the unmanipulated hematopoiesis (Rodriguez-Fraticelli et al., 2018; Yu et al., 2016). Yu et al. created a multi-fluorescent mouse model that enables molecular profiling and functional tracking of live cells *in vivo*. It was found that the endogenous HSC pool is composed of highly heterogenous multipotent clones with very different cell kinetics, where some HSC clones persist for long intervals while others are transient. It was also observed that most hematopoietic populations are mainly maintained by a few major HSC clones that are dominant over clones with smaller sizes. Furthermore, the stereotypical intraclonal behavior of HSCs was confirmed, i.e., daughter cells retain the characteristics of progenitor cells, such as lineage bias and proliferative potency. It was also shown that the sensitivity to the stress of inflammation or radiation is also a clone-specific feature. This result suggests again that the heterogeneous behavioral features of individual HSCs are established early in development and are maintained under different circumstances. In this work, gene expression and epigenetic features were linked to functional characteristics at a clonal level *in vivo*. It was found that epigenetic features, like DNA methylation and chromatin accessibility, dictate how HSC will behave in proliferation and lineage differentiation. In contrast, expression levels of lineage-specific genes were not associated with the HSC commitment behavior. This epigenetic memory is fixed before the experiment and persistent under homeostatic and stress conditions (Yu et al., 2016).

Studies at the clonal level using single-cell transplantation and a barcoding assay identified in the HSC compartment cells with long-term repopulating activity and self-renewal capacity but lineage commitment to megakaryocytes (Rodriguez-Fraticelli et al., 2018; Yamamoto et al., 2013, 2018). Interestingly, paired daughter cell assays combined with transplantation suggested that HSCs can differentiate directly to these megakaryocyte-restricted cells without passing lineage commitment, i.e., in a non-stepwise manner (Yamamoto et al., 2013).

With the additional introduction of other surface markers, a revised model of a classical hematopoietic tree emerged, including a direct shortcut into the megakaryocytic lineage (Haas et al., 2018; Wilkinson et al., 2020) (**Figure 1.4**). However, this hierarchical tree-like model assumes that all mature cells from peripheral blood are originated from a single HSC, which can self-renew and originate a pool of homogenous HSC or undergo lineage differentiation, producing distinct progenitor populations in a stepwise manner. And it cannot explain why the HSC pool consists of multiple HSC subtypes with the distinct molecular and functional features previously found. Therefore, a new model based om the famous Waddington's epigenetic landscape was proposed for hematopoiesis (Karamitros et al., 2018; Rodriguez-Fraticelli et al., 2018; Velten et al., 2017). Velten et al. integrated transcriptomics, flow cytometric, and functional data at the single-cell level to reveal that acquiring lineage-specific biases is a continuous process. During this process, each HSC gradually acquires transcriptomic lineage priming in a combination of multiple directions without passing through distinct hierarchically organized multi- and bi-potent progenitor populations. Thus, unilinear-restricted cells emerge directly from a continuum of low-primed undifferentiated hematopoietic stem and progenitor cells. This continuum contains multipotent progenitors and multilymphoid progenitors, constituting transitory states and not discrete progenitor cells. According to this model, in the Waddington's landscape HSCs reside in a flat valley at the top and the barriers that represent the acquisition of lineage biases, resulting from the distinct gene expression patterns that separate individual lineages, rise early, and expand gradually. Lineage commitment is established when barriers become insuperable (Velten et al., 2017) (**Figure 1.4**). This model more appropriately reflects

that transcriptional lineage programs are affected by the epigenetic landscape already present in the HSC compartment and are associated with the functional lineage biases of HSCs (Haas et al., 2018; Yokota, 2019). Interestingly, differentiation through the commitment barriers can be changed in very specific conditions. It was shown that the deletion of *Pax5* leads committed B cells (pre- / pro-B cells) to dedifferentiate back into uncommitted hematopoietic progenitors cells and then follow other differentiation paths and produce various myeloid and lymphoid *in vivo* and *in vitro* (Cobaleda et al., 2007; Mikkola et al., 2002; Nutt et al., 1999b; Rolink et al., 1999; Schaniel et al., 2002).

2. Objectives

Clonal monoallelic expression of autosomal genes has been well documented in studies performed in collections of clones expanded *in vitro* without undergoing differentiation or under limited differentiation. But its prevalence *in vivo* is controversial due to a lack of studies that reflects the technical challenge of tracking and isolating clonal cell populations *in vivo*. The murine hematopoietic system offers a unique opportunity to approach this biological question. In this system, it is possible to create clonal populations emerging from a single HSC and produce different lineages that undergo distinct specific genetic programs. In addition, there are numerous tools to label and then easily isolate by FACS different hematopoietic populations present in the bone marrow, secondary lymphoid organs, and blood. Moreover, it is known that the HSC compartment is heterogeneous. This heterogeneity is possibly the result of intrinsic epigenetic features established early in development that are clonally propagated to daughter cells, and may be responsible for lineage biases. Through single-HSC transplantation in mice, the main objective of this study is to perform for the first time a genome-wide transcriptome analysis of different lymphocyte populations emerging from a single HSC *in vivo*. This approach will evaluate whether genes with specific-allele expression patterns in the autosomes, which are established early in development, are clonally maintained through extensive cell division and differentiation. We will focus on the comparison of the same differentiated cell population (such as B and T populations) collected from different animals reconstituted with a single HSC (monoclonal mice) or several HSCs (the polyclonal controls). The cell population will be the constant in this setting, and the original HSC will be the variable.

As the entire project is based on the efficiency of mouse reconstitutions and the production of a truly monoclonal hematopoietic system, optimization and internal controls are essential. These are described in the first part of this work, which is divided in three parts. In **1). optimization of single-HSC reconstitutions**, the different cell sorting protocols are described and evaluated. In **2). multilineage and long-term of reconstitutions**, the mice reconstituted with single or multiple HSCs are characterized. And in **3). evaluating the quality of the collected samples** the monoclonality and purity of the samples are addressed. The second part is focused on the analysis of the genome-

wide transcriptome data. In **4). quality of RNA-seq and samples,** the quality of the transcriptome data and the samples are evaluated. In **5). identification of autosomal allele-specific expression**, the persistence of stable RMAE marks in HSC-derived lymphocytes through extensive cell proliferation and differentiation *in vivo* is addressed, RMAE is compared to XCI in terms of clonal stability, and the frequencies of RMAE in clones undergoing extensive differentiation *in vivo* and clones expanding *in vitro* without differentiation are compared. In **6). identification of XCI escapees**, a new method to study XCI and XCI escapees in genetically unmanipulated system *in vivo* is presented**.** Finally, in 7). **identification of genes with differential AI between B and T cells**, it is shown that these differences have a genetic basis.

3. Materials and methods

3.1. Animal breeding

All mice were bred and maintained at the specific pathogen-free animal facilities of the Instituto Gulbenkian de Ciência (IGC, Oeiras, Portugal). C57BL/6J-Ly5.1 (C57BL/6J strain carrying the pan-leukocyte marker Ly5.1), C57BL/6J-Ly5.2 (C57BL/6J strain carrying the pan-leukocyte marker Ly5.2), C57BL/6J-Ly5.2/ β-actin-GFP (C57BL/6J strain carrying the pan-leukocyte marker Ly5.2, and with GFP under the control of a chicken beta-actin promoter) and CAST/EiJ were originally received from The Jackson Laboratory (Bar Harbor, ME, USA). Animals used in reconstitution experiments were bred at our animal facility to generate female heterozygous F1 donor (CAST/EiJ x C57BL/6J-Ly5.2) and recipient (CAST/EiJ x C57BL/6J-Ly5.1) animals. Donor animals used in cell transfer experiments were <5 weeks old, and recipient animals were 5—16 weeks old. This research project was reviewed and approved by Órgão Responsável pelo Bem-Estar dos Animais (ORBEA) of IGC (PTDC/BEX-BCM/5900/2014 reference) that regulates the use of laboratory animals.

3.2. HSC isolation

The bone marrow was flushed out from the tibia and femur using a syringe and single-cell-suspended in FACS buffer (1x PBS, 2% FBS). The erythrocytes were lysed with red blood cell lysis buffer (RBC lysis buffer) (155 mM NH_4Cl, 10 mM $NaHCO_3$, 0.1 mM EDTA, pH 7.3) for 5 min and immediately rinsed and washed with FACS buffer. The cells were blocked with FcBlock (anti-CD16/32) at 4°C and washed. Enrichment for negative lineage cells was performed by incubating cell suspension with a cocktail of biotin-conjugated antibodies for surface markers of lineage-committed cells (**Table 3.1**) and, subsequently, lineage-marked cells were depleted using MACS Streptavidin MicroBeads (Miltenyi Biotec) for negative selection of lineage-positive cells by immunomagnetic separation using a MACS column (Miltenyi Biotec). Cells were further stained with PI and fluorophore-conjugated antibodies to isolate LT-HSCs (**Table 3.1**). During all washing steps, centrifugations were performed at 300 x g and 4°C. All staining incubations were carried out for 20 min at 4°C, except FITC-CD34, which was done for 90 min. LT-HSCs

were sorted on a FACSAriaII using the single-cell deposition unit into the individual wells of Terasaki plates (no. 452256, MicroWell 60-well MiniTray, Nunc Brand, Thermo Fisher Scientific Inc.) preloaded with 15 µL of FACS buffer. Each well was examined in a 4°C room using an inverted microscope, and only the wells with a single cell were used in the reconstitutions. Cells from wells were transferred to an individual 1.5 mL tube performing the washing of the wells. The 50—200 HSCs used to reconstitute polyclonal donor animals were sorted directly to the tube. To avoid cellular death, the cell suspension was always kept at 4°C starting from tibia and femur isolation until injection into recipient animals.

Table 3. 1. Antibodies used for staining lineage-committed cells and long-term hematopoietic stem cells according to the different protocols tested in this work.

Staining	Anti-	Conjugation
Lineage-committed cells	CD45R/B220	Biotin
	anti-CD19	
	CD11b/Mac1	
	Ly-76/Ter119	
	Ly6G/Gr1	
	CD3	
LT-HSCs protocol 1	CD34	FITC
	CD135	PE
LT-HSCs protocol 2	CD48	BV421
	CD150	PE
LT-HSCs protocol 3	CD34	A700
	CD135	PE
	CD49b	BV711

3.3. Animal reconstitutions

Recipient females (5–16 week-old) received sublethal whole-body g-irradiation with 600 cGy (Gammacell 2000 Mølsgaard Medical), 2–6 h before an intravenous retro-orbital injection with single-HSC or 50–200 HSCs. Recipient animals were analyzed routinely four weeks after injection and every two weeks for up to 12 weeks for chimeric cells in the peripheral blood. Blood samples were collected from the submandibular vein in EDTA. Erythrocytes were lysed using RBC lysis buffer for 5 min and immediately rinsed and washed with FACS buffer. Cells were further stained with PE-conjugated anti-Ly5.1

and FITC-conjugated anti-Ly5.2 antibodies washed and analyzed by FACSCanto or FACScan.

3.4. Processing of animal samples

Animals selected for subsequent analysis showing chimeric cells 12 weeks post-reconstitution were sacrificed and processed by removing thymi, spleens, and bone marrows. Single-cell suspensions from bone marrow were obtained as described above using a syringe and a 70-µm nylon mesh for the spleen and thymus. Erythrocytes were lysed with RBC lysis buffer for 5 min and immediately rinsed and washed with FACS buffer. Around 30% of cell suspension from bone marrow was saved for reconstitution of sublethally irradiated secondary recipient female mice, injected by intravenous retro-orbital administration, and analyzed for chimerism four weeks post-injection as described above. Different stainings with labeled antibodies were used to analyze and sort lymphoid populations in the spleen and thymus and myeloid population in bone marrow or spleen with FACSAriaII, after cell blocking with FcBlock (anti-CD16/32) and washing. Different antibodies (**Table 3.2**) were combined with PI in each single-cell reconstitution experiment. All FACS data were analyzed using the FlowJo program.

Table 3. 2. The different combinations of antibodies used in 16 single-cell reconstitution experiments.

Experiments	Anti-	Conjugation	Tissue
1-7	Ly5.1	APC-Cy7	Common
	Ly5.2	PE	
	CD19	PE-Cy7	Spleen
	IgM	APC	
	Mac1	BV786	
	CD4	PE-Cy7	Thymus
	CD8	BV605	
	B220	PE-Cy5	Bone marrow
	IgM	APC	
8-16	Ly5.1	FITC	Common
	Ly5.2	PE	
	CD19	PE-Cy7	Spleen
	IgM	APC	
	CD4	PE-Cy7	Thymus
	CD8	BV605	
	Mac1	BV786	Bone marrow

3.5. RNA and DNA extraction

After cell sorting, pellets were harvested by centrifugation and resuspended in 0.25 mL of cold TRIzol Reagent or 0.1 mL of Absolutely RNA Nanoprep Kit (Agilent #400753) lysis buffer. The suspension was homogenized by pipetting up and down several times to lyse the cells. Homogenized samples were stored at -80°C until isolation. RNA isolation with Absolutely RNA Nanoprep Kit was performed according to the manufacturer's protocols. Isolation with TRIzol followed the next steps. After incubation for 5 min for complete dissociation of the nucleoproteins complex, 0.1 mL of chloroform was added to the suspension, which was left 3 min for incubation and then centrifuged for 15 min at 12,000 x g at 4°C to separate the sample into a lower red phenol-chloroform, and interphase, and a colorless upper aqueous phase. The lower phase and interphase were saved for DNA extraction. The aqueous phase containing RNA was incubated with 10 µg of RNase-free glycogen for co-precipitation with RNA and 0.3 mL of isopropanol for 10 min at 4°C. The sample was centrifugated at 12,000 x g for 10 min at 4°C to recover the pellet, which was then resuspended in 0.7 mL of 75% ethanol, centrifugated at 7,500 x g for 5 min at 4°C and air-dried after discarding the supernatant. RNA was resuspended in 20 µL of RNase-free water and stored at -80ºC until processing.

0.2 mL of 100% ethanol was added to the lower phase and interphase for DNA extraction. The sample was incubated for 3 min and centrifugated at 2,000 x g at 4°C for 5 min to recover the pellet. This was resuspended in 0.5 mL of 0.1 M sodium citrate in 10% ethanol (pH 8.5), incubated for 30 min, and centrifugated at 2,000 x g at 4°C for washing. Next, the washed pellet was resuspended in 1 mL of 75% ethanol, incubated for 20 min, and centrifugated at 2,000 x g at 4°C. Finally, the supernatant was discarded, the pellet was air-dried, resuspended in 20 µL of H_2O, and stored at -20°C until sequencing.

3.6. Monoclonality screening

RNA was isolated from the same repopulated animals using sorted cell populations other than the sequenced ones to test for monoclonality before sequencing. According

to the manufacturer's recommendations, cDNA was prepared using SuperScript IV (ThermoFisher #18090050). The *Xist locus* was amplified in two individual reactions using two sets of primers to produce amplicons with two different SNPs: Fw1 5'agacgctttcctgaacccag with R1 5'aagatgctgcagtcaggc; and Fw2 5'ggagtgaagagtgctggagag with R2 5'gtcagtgccactattgcagc. PCR mix was performed with 0.75 units of GoTaq DNA polymerase (Promega #M3005), 1 x GoTaq reaction buffer, 0.2 mM of each dNTPs, 1.5 mM of $MgCl_2$, 1 µM of each primer, and 10—50 ng of RNA in a final volume of 10 µL. Amplicons were amplified using the following program: 5 min at 95°C, 45 cycles of 30 s at 95°C, 30 s at 60°C, and 25 s at 72°C, and a final elongation of 7 min at 72°C. The amplicons were separated in a 1.5 % agarose gel, purified with columns, and sequenced by Sanger sequencing with Fw1 or R2 primers.

3.7. cDNA library preparation and whole-transcriptome sequencing

Omega Bioservices, USA, performed cDNA library preparation and whole-transcriptome sequencing. According to the manufacturer's protocol, 2-3 RNA-sequencing libraries per RNA sample were prepared using SMART-Seq v4 Ultra Low Input RNA Kit (Clontech). Technical replicates of 10 ng of RNA were used as input. The RNA was primed by an oligo(dT) primer (3' SMART-Seq CDS Primer II A), and first-strand cDNA synthesis was performed at 42°C for 90 min and 70°C for ten min. The resulting cDNA was then amplified via PCR using the following program: 1 min at 95°C, eight cycles of 10 sec at 98°C, 30 s at 65°C, and 3 min at 68°C, and a final elongation of 10 min at 72°C. 15–200 pg full-length cDNA was tagged and fragmented by the Nextera XT transposome (Illumina) and amplified by PCR: 30 s at 95 °C, 12 cycles of 10 s at 95 °C, 30 s at 55 °C, and 30 s at 72 °C, then 5 min at 72 °C. Mag-Bind RxnPure Plus magnetic beads (Omega Bio-tek) were used to purify the library and provide a size-selection step. The libraries were then pooled in equimolar concentrations and sequenced on Illumina HiSeq 2500 machine (150 bp, paired-end).

3.8. DNA library preparation and whole-exome sequencing

DNA samples (E6.2-B220^{+}IgM^{+} from bone marrow, E6.43-CD4^{+}CD8^{-} from the thymus, and E15.10-CD4^{+}CD8^{-} from thymus) were used for whole-exome sequencing (WES). Novogene, UK, performed DNA library preparation and whole-exome sequencing using Agilent SureSelect Mouse All ExonV6 kit (Agilent Technologies) following the manufacturer's recommendations, and x index codes were added to attribute sequences to each sample. The genomic DNA samples were randomly fragmented by sonication (Covaris) to the size of 180–280 bp fragments. The remaining overhangs were converted into blunt ends via exonuclease/polymerase activities. Adapter oligonucleotides were ligated after adenylation of 3’ ends of DNA fragments. DNA fragments with ligated adapter molecules on both ends were selectively enriched in a PCR reaction. The libraries were hybridized with biotin-labeled probes, and magnetic beads with streptomycin were used to capture the exons. After washing beads and digesting the probes, the captured libraries were enriched in a PCR reaction to add index tags. The products were purified with the AMPure XP system (Beckman Coulter). DNA libraries were sequenced on an Illumina platform (150 bp, paired-end). Read alignment and allele counts were based on the ASEReadCounter* pipeline. Genes with total allelic counts of <10 and genes with nominal AI >0.75 or <0.25 were excluded.

3.9. VDJ clonotypes

Immunoglobulin rearrangements were detected by aligning RNA-seq raw data with reference germline V, D, J, and C gene sequences, and assembled into clonotypes with MiXCR-3.0.12 15 (Bolotin et al., 2015, 2017).

3.10. Allele-specific gene expression analysis from RNA-seq

RNA-seq data analysis for allelic imbalance estimation followed the ASEReadCounter* tool adapted from the GATK pipeline (Castel et al., 2015) for the pre-processing read alignment steps up to allele counts, and the statistical R package Qllelic.v0.3.2 for

calculation of the QCC (quality control correction for technical replicates) and estimation of confidence intervals for differential allelic imbalance analysis (Mendelevich et al., 2021). RNA-seq reads were trimmed from nextera adapters with cutadapt.v.1.14 using the wrapper trim_galore. Sequencing reads were aligned with the reference pseudogenome (maternal) and imputed (paternal) with the STAR aligner v.2.5.4a, with default filtering parameters and accepting only uniquely aligned reads. Samtools mpileup (v.1.3.1) was used to estimate allele-specific coverage over SNPs. Point estimates of allelic imbalance for a gene were obtained as the ratio of maternal allele counts over total allelic gene counts, excluding genes with <10 counts. Pairwise comparison of differential AI was performed using the binomial test a with QCC correction. Gene abundance counts were obtained with featureCounts from the same bam files generated with the ASEReadCounter* alignment pipeline, and abundance was estimated as TMM (trimmed mean of M) - normalized counts with edgeR (Robinson and Oshlack, 2010). Differential gene expression for B and T samples was performed with the same edgeR tool.

3.11. t-distributed stochastic neighbor embedding (t-SNE) analysis

t-SNE analysis of AI values was performed with the tsen function from the M3C algorithm (John et al., 2020).

3.12. Abelson clones

The v-Abl pro-B clonal cell lines Abl.1, Abl.2, Abl.3, and Abl.4 were derived previously from 129S1/SvImJ x Cast/EiJ F1 female mice by expansion of FACS-sorted single cells after immortalization (Zwemer et al., 2012). Immortalized B-cell clonal lines were cultured in Roswell Park Memorial Institute (RPMI) medium (Gibco), containing 15% FBS (Sigma), 1x L-Glutamine (Gibco), 1x Penicillin/Streptomycin (Gibco), and 0.1% β-mercaptoethanol (Sigma). The culture medium also contained 1% DMSO (because these cells were the control set in an experiment using a drug dissolved in DMSO). On day 2 of

the culture, $5x10^6$ live cells were collected after sucrose gradient centrifugation (Histopaque-1077, Sigma, Cat 10771). RNA was extracted from $2x10^5$ cells using a magnetic field bead-based protocol using Sera-Mag SpeedBeadsTM (GE Healthcare). According to manufacturers ' instructions, two libraries were prepared per clone using the SMARTseqv4 kit (Clontech), starting with 10 ng input RNA for each library. The Abl.1 clone was sequenced on an Illumina NextSeq 500 machine (75 bp, single-end); clones Abl.2, Abl.3, and Abl.4 were sequenced on an Illumina HiSeq 4000 machine (150 bp, paired-end). RNA-seq data analysis followed the same pipeline as for HSC-derived clones *in vivo*, except the maternal reference genome, which was 129S1. These data were originally generated for the work described in Gupta et al. (Gupta et al., 2021).

After sucrose gradient collection, the remaining cells were washed with 1x PBS and frozen on dry ice for genomic DNA extraction by GenElute Kit (Sigma, #G1N10-1KT). LC Sciences (TX, USA) performed library preparation, QC, and whole-exome sequencing (50x). SureSelect (Agilent Technologies) was used for exome capture following the manufacturer's recommendations. A Hiseq X Ten sequencing instrument (Illumina) (150 bp, paired-end) generated reads. Read alignment and allele counts were based on the pipeline as RNA-seq of Abelson clones, genes with total allelic counts of <10, and those with nominal AI >0.7 or <0.3 were excluded (Gupta et al., 2021).

3.13. XCI escapees

X- linked genes were considered XCI escapees if significant expression from the inactive X chromosome was identified in each single-HSC derived sample by comparing the allelic ratio value for a given gene with a threshold value calculated for each sample as the median of the allelic imbalance distribution for all genes on that sample (to account for potential contamination from recipient cells) ± 0.1. The comparisons were performed by applying the binomial test with quality control correction for technical replicates (QCC) (Mendelevich et al., 2021). To consider a gene as an escapee, we defined three criteria: 1) only samples with expression higher than 10 TMM-normalized counts (Robinson and Oshlack, 2010) were considered; 2) the median of AI in the control

samples (polyclonal and unmanipulated samples) was balanced (0.5±0.2); 3) and AI was above the monoclonal sample threshold in at least two samples from the same tissue (B or T cells) or above in at least one B cell sample and at least one T cell sample (**Figure 4.21**).

3.14. Annotation of XCI escapees along the X chromosome

Annotation of XCI escapees along the X chromosome was carried out with the karyoploteR package (Gel and Serra, 2017).

3.15. Enrichment analysis

Statistical enrichment analysis of genes with differential AI between B and T cells was performed with g: Profiler (Peterson et al., 2020; Raudvere et al., 2019) against all known genes in the mouse genome, using a hypergeometric test with the default correction for multiple testing (set counts and size). As a control group with the same dimension (146 genes), a set of randomly selected genes without differential AI between B and T cells was used.

3.16. Statistical analysis

The difference between the AI point estimates of two clones, or the difference of point estimate and a threshold for XCI escapees, was accepted as significant after accounting for the experiment-specific overdispersion of 2-3 technical replicates with QCC using the R package Qllelic.v0.3.2 (Mendelevich et al., 2021). All QCC corrected values incorporate Bonferroni correction to account for multiple hypothesis testing. Statistical analyses in the study were conducted in R.

3.17. Data availability

The entire set of HSC next-generation sequencing (NGS) raw data (RNA-seq and whole-exome sequencing) and processed counts files have been deposited to the NCBI's Gene Expression Omnibus database with the series accession number [GEO: GSE174040]. Abelson clones RNA-seq (Gupta et al., 2021) data have been deposited with the series accession number [GEO: GSE144007].

4. Results

4.1. Hematopoietic stem cell reconstitutions

Vasco M. Barreto contributed to the single-cell reconstitution experiments, data analysis, and graphical visualization.

4.1.1. Introduction

We took advantage of HSCs, which have been studied for a long time and in an exhaustive way. They are more well-characterized than other types of adult mammalian stem cells. Another convenience of HSCs is that it is much easier to transplant them than other stem cells. It is possible to remove HSCs from their niches in a donor mouse and inject them into a recipient animal's circulatory blood system, which was previously preconditioned to lack a functional endogenous hematopoietic system. These transplanted cells can survive, find their way back to their niches, and retain HSC potential, thus repopulating the bone marrow and producing all hematopoietic lineages. The single-cell transplantation assay has been used as a gold standard to define the functional capacity of HSCs, which are characterized by multipotency and sustained self-renewal (Cheng et al., 2020; Mayle et al., 2013; Schroeder, 2010). Multipotency is described by the capacity of HSCs to reproduce the entire hematopoietic system, which consists of differentiation into myeloid and lymphoid cell populations. The self-renewal capacity is evaluated by the reconstitution kinetics. Initially established multilineage populations from an injected single HSC should be stable and persist at least four months after transplantation. Additionally, serial transplantation, in which donor cells are collected from primary recipients and injected into secondary recipients, is used to demonstrate that the original donor HSC can reconstitute the primary recipient system and its progeny can successfully reconstitute the secondary recipient, indicating that HSC has long-term potent self-renewal capacity (Dykstra et al., 2007; Kiel et al., 2005; Wilkinson et al., 2020). In contrast, short-term HSCs can reconstitute primary recipients for four months, but they lack potent self-renewal capacity and cannot reconstitute secondary recipients (Wilkinson et al., 2020). Another advantage of the single-cell transplantation assay is that all progeny detected at any time can be assigned to the

same original donor HSC and enable the study of hematopoietic cell populations at the clonal level *in vivo* (Dykstra et al., 2007).

The produced clonal hematopoietic populations will be analyzed through whole-transcriptome sequencing to study their allele-specific expression. This analysis is necessary to distinguish between the expression of maternal and paternal alleles in cells. For this purpose, two murine strains (Cast/Ei (CAST) and C57BL/6 (B6)) that are genetically very distant were crossed to obtain a highly heterozygous F1 progeny (Frazer et al., 2007). These mice have a high SNP frequency and, consequently, around 83% of the genes can be evaluated by the transcriptome analysis (Eckersley-Maslin et al., 2014). Additionally, to evaluate the chimerism level in blood or primary and secondary hematopoietic organs, we used congenic B6 strains that carry different alleles of *Ptprc*. This gene encodes for two different isoforms of the pan-leukocyte surface marker Ly5, known as Ly5.1 and Ly5.2. As the name indicates, this marker is expressed on the surface of leukocytes (including myeloid and lymphoid populations). Using monoclonal antibodies for the two isoforms and flow cytometry, it is possible to distinguish cells derived from the single donor cell and recipient cells in a simple and consistent way (Wilkinson et al., 2020).

As the success of this project depends on the establishment of the monoclonal hematopoietic system, we used only female animals for reconstitutions. This choice allows us to take advantage of XCI and evaluate *a posteriori* if recipient animals were reconstituted with one or very few cells as opposed to many HSCs. To guarantee dosage compensation of X-linked genes between XX females and XY males, one of the X chromosomes is randomly inactivated early in the development of the female mammalian embryo. The expression of *Xist* induces the inactivation of the X chromosome from which this non-coding RNA is stochastically transcribed (Maduro et al., 2016). This choice is clonally propagated and leads to mosaicism. One-half of the cells have inactivated the paternal X chromosome, and the other half have inactivated the maternal X chromosome. By injecting a single HSC from a female that has previously inactivated one of the two X chromosomes, we should observe the emerging hematopoietic cells all with the same inactivated X chromosome. On the other hand, if

more than one HSC is injected to reconstitute the recipient animal, in most cases we should detect a mixture of cells with the paternal and maternal inactivate X chromosomes (**Figure 1.1**). Given the high density of SNPs present in heterozygous F1 hybrid mice, allele-specific expression of *Xist* by Sanger sequencing and later of X-linked genes by whole-transcriptome sequencing will be applied to confirm the monoclonality of the collected samples.

4.1.2. Optimization of single-HSC reconstitutions

To study stable allele-specific transcriptional states, we introduced a single HSC from a donor female mouse into a sub-lethally irradiated recipient female animal. Heterozygous F1 hybrid donor mice were produced by crossing CAST x $B6^{Ly5.1/Ly5.1}$. In parallel, F1 hybrid recipient mice were obtained by crossing CAST x $B6^{Ly5.2/Ly5.2}$. As mentioned before, the pan-leukocyte surface markers Ly5.1 and Ly5.2 are labeled by two different antibodies (which do not label the leukocytes from CAST; **Figure 4.1 A**). The injected single cell was left to expand and differentiate inside the donor for at least 12 weeks to produce clonal multilineage hematopoietic cell populations (monoclonal animals). Additionally, 50–200 HSCs were also transplanted per animal to generate oligoclonal or polyclonal control populations (oligoclonal or polyclonal animals). At the end of 12 weeks after injection, animals with chimerism levels in the peripheral hematopoietic system higher than 1% (monitored by blood flow cytometric analysis) were used to collect different cell populations from different organs, and bone marrow cells were used for secondary transplantation to confirm the self-renewal potential of injected HSCs (**Figure 4.1 B**).

The first implemented step in this work was the optimization of protocols for the isolation of highly pure HSCs. The HSC population is highly heterogeneous, with different subpopulations belonging to LT-HSC, ST-HSC, and even IT-HSC (Benveniste et al., 2010). There is no single marker to discriminate LT-HSCs from other hematopoietic cells in the bone marrow. Several protocols based on multiparameter flow cytometry were developed to isolate highly pure HSCs. The conventional strategy to obtain HSCs starts

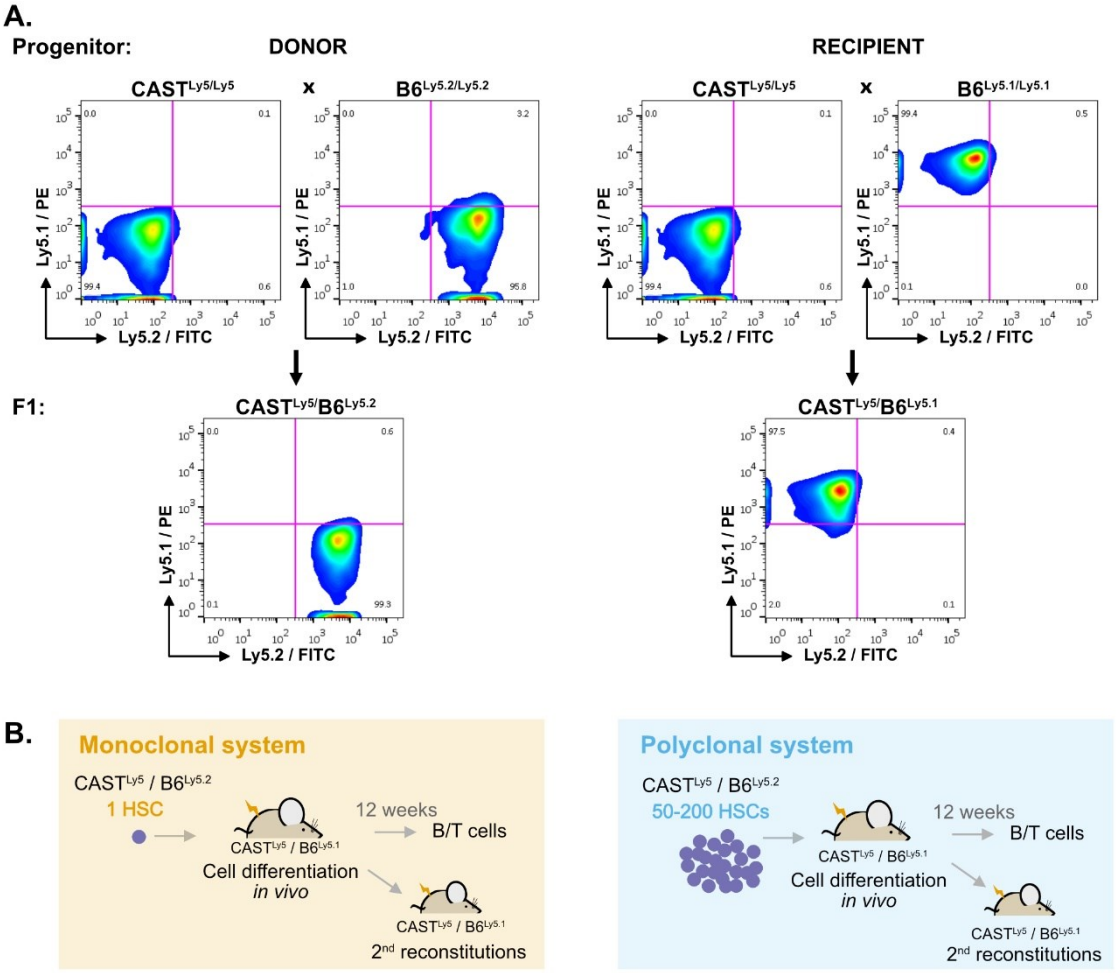


Figure 4. 1. Strategy to produce the monoclonal hematopoietic system *in vivo*. (A) Ly5.1 and Ly5.2 pan-leukocytic markers distinguish recipient from donor cells in reconstituted animals, respectively. Ly5.1 and Ly5.2 do not label the CAST progenitor line. When CAST is crossed with $B6^{Ly5.1/Ly5.1}$ and $B6^{Ly5.2/Ly5.2}$ to produce the recipient and donor F1 animals, respectively, the recipient and donor cells are distinguishable using these two markers. Blood samples of progenitor and descendants (F1) were lysed for red cells, stained with FITC-conjugated anti-Ly5.2 and PE-conjugated anti-Ly5.1, and analyzed using FACSCanto. (B) Schematic representation of monoclonal and polyclonal hematopoietic system establishment *in vivo*. A single hematopoietic stem cell (HSC) or 50–200 HSCs were injected into sub-lethally irradiated recipient mice to generate a monoclonal or polyclonal hematopoietic system. Different donor mice were used in each experiment. Secondary reconstitutions and isolation of B/T cell populations were performed after 12 weeks of cell differentiation *in vivo*.

with enrichment of the bone marrow population with stem cells. A cocktail with a combination of different markers against differentiated cells is used to deplete lineage-positive blood cells, thus selecting the negative lineage population (Lin^-). In addition, positive selection for Sca-1 and cKit markers, known to be expressed on the surface of HSCs, is used to obtain the LSK ($Lin^-Sca\text{-}1^+cKit^+$) population, which contains LT-HSCs but still has ST-HSCs and MPPs (together known as hematopoietic stem and progenitor cells (HSPCs)). Several additional selection strategies using specific surface markers are applied to isolate pure LT-HSCs (Mayle et al., 2013). We tested three different protocols for HSC isolation and two approaches for single-cell reconstitution (**Figure 4.2 A**). All three protocols were based on the first gating to select the LSK population for HSC enrichment and then for LT-HSCs: **1).** $CD34^-CD135^-$ (Adolfsson et al., 2001); **2).** $CD48^-CD150^+$ (Kiel et al., 2005); and **3).** $CD34^-CD135^-CD49b^-$ (Benveniste et al., 2010). We also tested two approaches with protocol 2 ($CD48^-CD150^+$) based on SLAM family markers. In the second approach (protocol 2.2), we injected 2 HSCs in the same donor animal to improve the percentage of reconstituted animals, expecting to double the chance of reconstitutions. However, since we need to study clonally expanded blood cells, and to distinguish between 2 injected HSCs, we used one HSC from donor F1 hybrid mouse derived from crossing CAST x B6 $^{Ly5.2/Ly5.2}$ / $^{\beta\text{-actin-GFP}/\ \beta\text{-actin-GFP}}$. These mice present widespread GFP fluorescence, enabling the discrimination from the second injected HSC without fluorescence (**Figure 4.2 A**). The HSCs were single-cell sorted by flow cytometry into Terasaki plates, and each well was confirmed to contain only one cell under a microscope inside a cold room to avoid cell death (**Figure 4.2 B**). After that, cells were introduced intravenously by retro-orbital injection into sublethally irradiated recipient mice and allowed to expand *in vivo*. A different donor was used for each experiment.

The contribution of the donor and recipient HSCs to the peripheral hematopoietic system was analyzed by staining leukocytes for the donor cells using the Ly5.2 antibody. The evaluation of chimerism was performed starting from week four post-injection, as a single HSC needs several self-renewal and differentiation steps to produce mature progeny and become detectable in the blood (Muller-Sieburg et al., 2002). Donor cell estimation was performed every 14 days for eight weeks. In total, 16 experiments were

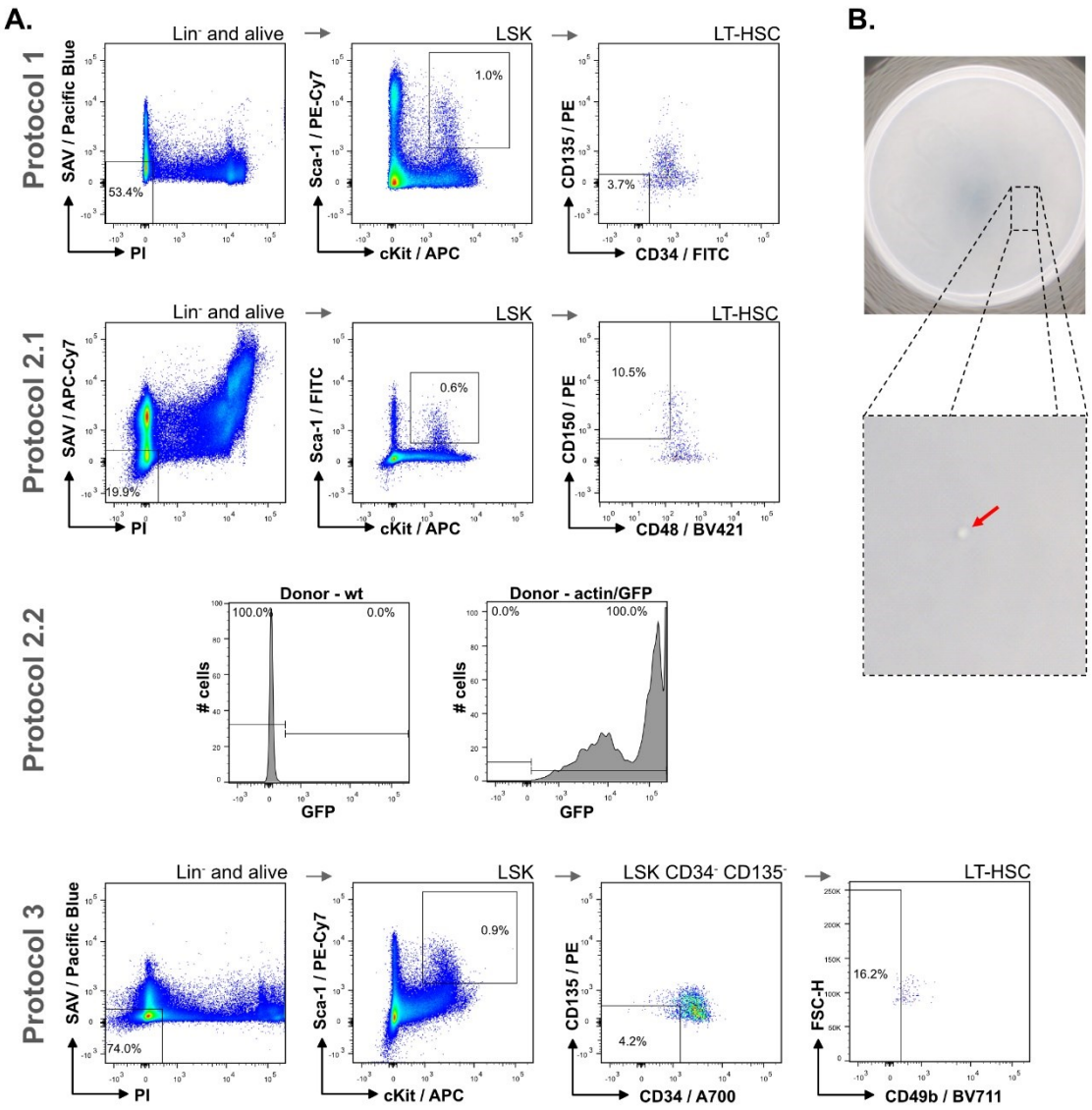


Figure 4. 2. Isolation of pure long-term HSC (LT-HSC) population. (A) The different protocols used to separate LT-HSC from short-term HSC and progenitor cells. All protocols included the first step of lineage-marked cells depletion using MACS Streptavidin MicroBeads. For this, the bone marrow cells of an F1 CAST$^{Ly5/Ly5}$ x B6$^{Ly5.2/Ly5.2}$ (protocols 1,2.1, 2.2 and 3) or B6 $^{Ly5.2/Ly5.2}$ / β-actin-GFP/ β-actin-GFP (protocol 2.2) mouse were stained with a cocktail of biotin-conjugated antibodies for surface markers of lineage-committed cells (anti-B220, anti-CD19, anti-Mac1, anti-Ter119, anti-Gr1, and anti-CD3). After depletion, cells were stained with fluorophore-conjugated antibodies according to each protocol. Protocol 1: APC-conjugated anti-c-Kit, PE-Cy7-conjugated anti-Sca-1, PE-conjugated anti-CD34, FITC-conjugated anti-CD135, Streptavidin/Pacific-blue (SAV/PB), and PI, and sorted on a FACSAria. The cells were gated for PI$^-$ / SAV$^-$ to exclude dead cells and any remaining lineage-positive cells, then for c-Kit$^+$/Sca-1$^+$ to obtain Lin$^-$Sca$^+$c-Kit$^+$ (LSK) cells, and finally gated for CD34$^-$/CD135$^-$ to obtain LT-HSCs. Protocol 2.1: APC-conjugated anti-c-Kit, FITC-conjugated anti-Sca-1, BV421-conjugated anti-CD48, PE-conjugated anti-CD150, Streptavidin/APC-Cy7 (SAV/APC-Cy7), and PI, and sorted on a

FACSAria. The cells were gated for PI^- / SAV^- to exclude dead cells and any remaining lineage-positive cells, then for c-Kit^+/$Sca\text{-}1^+$ to obtain $Lin^-Sca^+c\text{-}Kit^+$ (LSK) cells, and finally gated for $CD48^-/CD150^+$ to obtain LT-HSCs. Protocol 2.2: APC-conjugated anti-c-Kit, PE-Cy7-conjugated anti-Sca-1, BV421-conjugated anti-CD48, PE-conjugated anti-CD150, Streptavidin/APC-Cy7 (SAV/APC-Cy7), and PI, and sorted on a FACSAria. The cells were gated for PI^- / SAV^- to exclude dead cells and any remaining lineage-positive cells, then for c-Kit^+/$Sca\text{-}1^+$ to obtain $Lin^-Sca^+c\text{-}Kit^+$ (LSK) cells, and finally gated for $CD48^-/CD150^+$ to obtain LT-HSCs. GFP^+ and GFP^- donor bone marrow cells were stained separately with the same antibodies for this approach. Then each cell population was individually single-cell sorted. Protocol 3: APC-conjugated anti-c-Kit, PE-Cy7-conjugated anti-Sca-1, A700-conjugated anti-CD34, PE-conjugated anti-CD135, Streptavidin/Pacific-blue (SAV/PB), and PI, and sorted on a FACSAria. The cells were gated for PI^- / SAV^- to exclude dead cells and any remaining lineage-positive cells, then for c-Kit^+/$Sca\text{-}1^+$ to obtain $Lin^-Sca^+c\text{-}Kit^+$ (LSK) cells, and finally gated for $CD34^-/CD135^-$ and $CD49b^-$ to obtain LT-HSCs. (B) After single-cell sorting into Terasaki pates, each well was confirmed under the microscope to contain only one HSC.

carried out. In each experiment, different HSCs were injected (**Table 4.1**). Experiments were executed as long as recipient animals of the appropriate age were available (without waiting for the accomplishment of the previous experiment). These mice eventually became unavailable because we were never able to breed the CAST line (CAST x CAST) and we eventually ran out of fertile CAST mice to breed with B6 animals. From the 16 performed experiments (**Table 4.1**), in the groups of animals injected a with a single and 50–200 cells the average percentages of reconstituted animals were 7.7% (35/453) and 76.9% (30/39), respectively. The level of chimerism in the blood of the monoclonal animals (reconstituted with single-cell) ranged from 1% to 44%, whereas the levels of polyclonal animals (reconstituted with 50–200 cells) varied from 2% to 88% (**Figure 4.3**). This result is expected, as animals reconstituted with more cells have a higher probability of developing donor hematopoietic systems since more cells are likely to survive, find the way back to niche, expand, and develop multilineage populations.

The first four experiments showed that HSCs isolated with protocol 2.1 (LSK $CD48^-$ $CD150^+$) produced more reconstituted animals, probably because this protocol is less time-consuming, and more HSCs can survive. From experiment 5, we applied the second strategy for protocol 2 (protocol 2.2), but with the introduction of two HSCs (GFP^+ and

GFP$^-$), we did not observe higher numbers of animals with chimerism. Protocol 3 (experiment 7) also did not produce more reconstituted animals. We, therefore, observed that protocol 2.1 seems to be more robust, and from experiment 8 onwards, we used only protocol 2.1. In experiments 10–12, we used older animals which could have negatively influenced the reconstitutions, even when a pool of HSCs was used - the case of polyclonal animals in experiments 11 and 12. We did not use animals from experiments where polyclonal recipient animals were not reconstituted (experiments 2, 11, and 12). Only animals (monoclonal and polyclonal) with levels of Ly5.2$^+$ donor cells in the blood higher than 1% and able to reconstitute secondary recipients were considered.

Table 4. 1. Percentages of reconstituted animals at 12 weeks post-injection and the protocol used for the 16 experiments. Exp., experiment; reconst., reconstituted; recip., recipient.

		Monoclonal (1 HSC)			Polyclonal (50–200 HSCs)		
Exp.	**Protocol**	**# injected animals**	**# reconst. 1st recip.**	**% of reconst. animals**	**# injected animals**	**# reconst. 1st recip.**	**% of reconst. animals**
1	2.1	29	2	7	4	4	100
2	1	28	1	4	2	0	0
3	2.1	26	4	15	2	2	100
4	1	21	2	10	2	2	100
5	2.2	54	3	6	3	2	66
6	2.1	58	11	19	4	4	100
7	3	21	1	5	4	4	100
8	2.1	18	1	6	2	1	50
9	2.1	26	1	4	2	1	50
10	2.1	15	0	0	2	2	100
11	2.1	15	0	0	2	0	0
12	2.1	16	0	0	2	0	0
13	2.1	34	2	6	2	2	100
14	2.1	36	0	0	2	2	100
15	2.1	26	4	15	2	2	100
16	2.1	30	3	10	2	2	100
mean				7.7			76.9

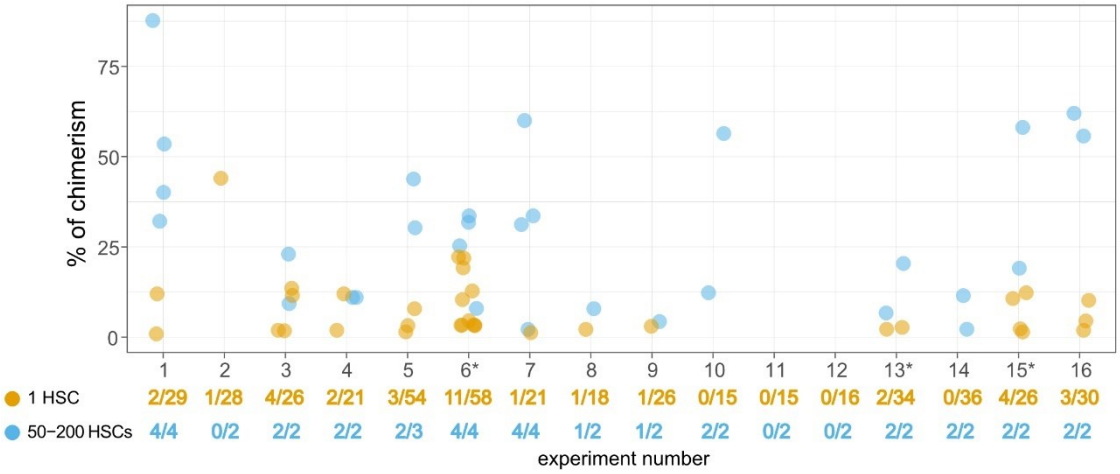


Figure 4. 3. Levels of chimerism for the different experiments. Percentages of chimerism identified in the blood of reconstituted animals for 16 experiments at 12 weeks post-injection (orange dots, monoclonal animals; blue dots, polyclonal animals; fraction, number of animals with chimerism/number of injected animals; asterisk, experiments used for further RNA-seq analysis). An animal was considered reconstituted if the chimerism percentage was above 1%.

4.1.3. Multilineage and long-term reconstitutions

The classical criteria to evaluate the purity of HSC is through long-term reconstitution. Three parameters analyzed the quality of these reconstitutions. First, we studied the ongoing production of leukocytes in the peripheral hematopoietic system for 12 weeks (**Figure 4.4 A**), showing that the initial HSC can differentiate and feed the hematopoietic system for at least 12 weeks. After this time, monoclonal and polyclonal animals with chimerism higher than 1% in the blood were sacrificed and used to isolate different HSC-derived donor cell populations (including IgM^+CD19^+, $CD4^+CD8^+$, and $Mac1^+$). The presence of lymphoid and myeloid cell populations in the primary (bone marrow and thymus) or secondary (spleen) hematopoietic organs (**Figure 4.4 B and Table 4.2**) confirms that the original HSC can differentiate into different hematopoietic cells, producing multilineages. The potential of the donor HSC self-renewal was confirmed by secondary recipient reconstitutions produced by injection of bone marrow cells from primary recipient animals (**Figure 4.4 C and Table 4.2**), which indicates that injected HSC was able to expand and produce more daughter HSCs. Thus, the HSCs used for single-cell injections and reconstitutions meet the definition of LT-HSCs (Dykstra et al., 2007; Kiel et al., 2005; Wilkinson et al., 2020).

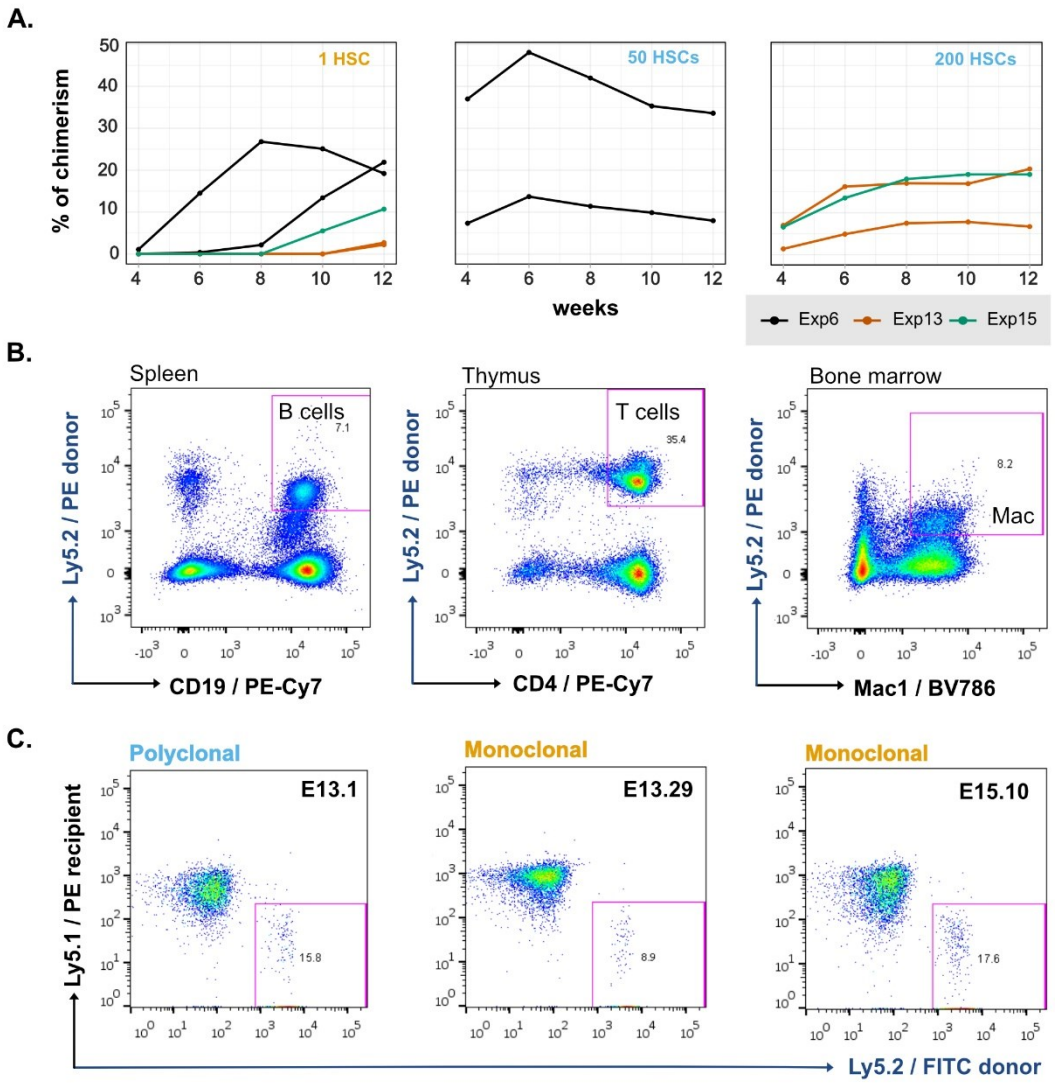


Figure 4. 4. A single HSC gives rise to myeloid and lymphoid cells in the blood with long-term reconstitution. (A) Evolution of donor-derived cell population percentages over time in the peripheral blood of the recipient animals. After blood collection, red cells were lysed, stained for Ly5.2 cells, and analyzed in a FACSCanto or FACScan instrument. (B) A single donor HSC differentiates into lymphoid and myeloid hematopoietic populations *in vivo*. Cells from different hematopoietic organs of recipient animals were isolated, stained, and gated on PI^-, FITC anti-$Ly5.1^+$, PE anti-$Ly5.2^-$ and PE-Cy7 anti-$CD19^+$ (spleen), PE-Cy7 anti-$CD4^+$ (thymus), or BV786 anti-$Mac1^+$ (bone marrow). (C) A single donor HSC repopulates secondary recipients. Plots of secondary reconstitutions four weeks post-reconstitution with bone marrow cells isolated from polyclonal and monoclonal primary reconstituted animals are represented. Blood samples of secondary reconstituted mice were lysed for red cells, stained with FITC-conjugated anti-Ly5.2 for donor cells, and PE-conjugated anti-Ly5.1 for recipient cells and analyzed using FACSCanto. Representative plots are shown.

Table 4. 2. A summary of the reconstituted animals used for whole-transcriptome sequencing. Only animals that could produce long-term multilineage reconstitutions reconstitute 2nd recipients and originate samples that passed the monoclonality and purity assays were used for whole-transcriptome analysis.

Experiment	Clonality	Sample	2nd reconstitution	Monoclonality assay (SNP *Xist*)	Purity assay (SNP *Ly5*)	Cell population	% of cell population	% of post-sort	RIN	Sequenced
6	Polyclonal	E6.1	y	Both	Ly5.2	IgM CD19	30.8	96.6	9.9	y
						CD4 CD8	3.0	97.5	0.0	n
						Mac1	14.5	97.8	Na	n
		E6.2	y	Both	Ly5.2	IgM CD19	9.7	92.9	8.0	y
						CD4 CD8	4.1	99.3	Na	n
						Mac1	3.5	94.2	Na	n
	Monoclonal	E6.42	y	B6	Ly5.2	IgM CD19	21.2	96.8	9.9	y
						CD4 CD8	16.9	99.6	0.0	n
						Mac1	3.2	94.9	Na	n
		E6.43	y	B6	Ly5.2	IgM CD19	23.9	94.2	6.6	y
						CD4 CD8	72.8	99.4	Na	n
						Mac1	35.0	95.7	Na	n
13	Polyclonal	E13.1	y	Both	Ly5.2	IgM CD19	21.0	96.8	10.0	y
						CD4 CD8	26.0	100.0	8.3	y
						Mac1	16.5	99.4	Na	n
		E13.2	y	Both	Ly5.2	IgM CD19	6.9	98.0	9.7	y
						CD4 CD8	19.4	99.3	10.0	y
						Mac1	9.2	96.2	Na	n
	Monoclonal	E13.24	y	CAST	Ly5.2	IgM CD19	2.5	96.4	6.9	y
						CD4 CD8	2.3	99.0	9.8	y
						Mac1	1.9	92.6	Na	n
		E13.29	y	CAST	Ly5.2	IgM CD19	2.5	99.2	9.5	y
						CD4 CD8	7.5	97.1	10.0	y
						Mac1	2.2	96.4	Na	n
15	Polyclonal	E15.2	y	Both	Ly5.2	IgM CD19	18.1	98.6	9.9	y
						CD4 CD8	15.3	100.0	0.0	n
						Mac1	21.1	99.0	Na	n
	Monoclonal	E15.10	y	CAST	Ly5.2	IgM CD19	10.6	99.2	9.2	y
						CD4 CD8	36.0	98.9	9.9	n
						Mac1	14.2	96.6	Na	n
Na	Non clonal	Control	Na	Na	Na	IgM CD19	Na	Na	10.0	y
			Na	Na	Na	CD4 CD8	Na	Na	9.6	y

4.1.4. Evaluating the quality of the collected samples

As mentioned before, this project's goal depends on producing monoclonal hematopoietic systems. Therefore, one of the first important steps for the quality of the collected samples is the isolation of pure donor cell populations grown in the recipient system. After sorting by FACS of hematopoietic cells, samples were checked for purity by re-running the same cells again through the FACS machine to calculate the percentage of cells that falls in the same gating Ly5.2-donor cell population used for isolation (**Figure 4.5 and Table 4.2**). Usually, the frequency of post-sorted cells is lower than 100%. This can be due to some loss of cell viability or fluorescence of cell surface markers. Therefore, we used cell samples with purity higher than 92% for further whole-transcriptome sequencing.

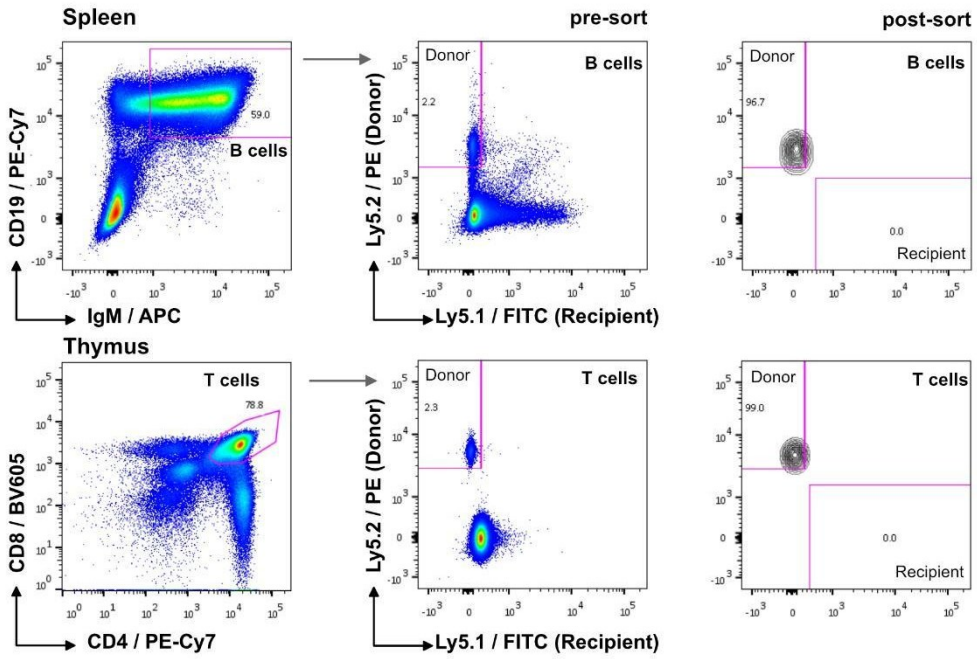


Figure 4. 5. Representative plots of pre-sorted and post-sorted B and T-cell populations of an animal reconstituted with a single HSC. Cells from the spleen and thymus of the recipient animal were isolated, stained for B-cell markers with PE anti-Ly5.2, FITC anti-Ly5.1, and PE-Cy7 anti-CD19 and APC anti-IgM (splenocytes), or T-cell markers with PE-Cy7 anti-CD4 and BV605 anti-CD8 (thymocytes), and sorted on a FACSAria. The cells were gated for PI^- to exclude dead cells and $CD19^+/IgM^+$ to select B-cells or for $CD4^+/CD8^+$ to select T-cells $Ly5.2^+/Ly5.1^-$ to obtain pure donor cells. The purity of sorted cells was assessed by analyzing 150–250 sorted cells.

The purity of the sorted donor cells was also studied by Sanger sequencing of *Ly5* cDNA from isolated RNA (**Figure 4.6 and Table 4.2**). We identified 3 SNPs present in the *Ly5* transcript that differ from *Ly5.1* (B6 recipient) and *Ly5.2* (B6 donor) isoforms. At the same time, these SNPs are identical for *Ly5.2* and *Ly5* (CAST) isoforms, making possible the distinction between donor and recipient cells. All analyzed samples (polyclonal and monoclonal) showed only SNPs corresponding to *Ly5.2* and *Ly5* (CAST) isoforms, which means that the collected samples were not heavily contaminated with recipient cells. However, this technique is not quantitative, and subsequently the percentage of contamination with recipient cells was calculated from RNA-seq data (see below).

SNPs (Ly5/Ptprc)		Donor ($CAST^{Ly5/Ly5}$ x $B6^{Ly5.2/Ly5.2}$)		Recipient ($CAST^{Ly5/Ly5}$ x $B6^{Ly5.1/Ly5.1}$)	
Position_id		$CAST^{Ly5}$	$B6^{Ly5.2}$	$CAST^{Ly5}$	$B6^{Ly5.1}$
1	**1_138107837**_rs33604282	T	T	T	C
	-	-	-	-	-
2	**1_138107824**_rs260353747	G	G	G	T
3	**1_138107823**_rs33604283	T	T	T	C

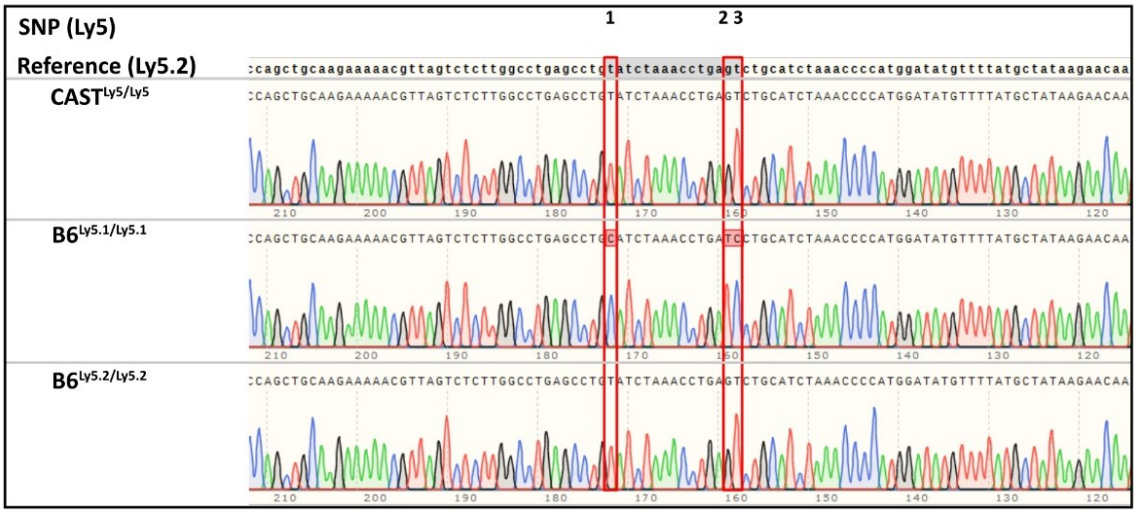


Figure 4. 6. Estimation of donor population contamination with recipient cells using Sanger sequencing. Identification and cDNA Sanger sequencing focus on three different SNPs for the *Ly5* gene, distinguishing two pan-leukocytic markers, Ly5.1 and Ly5.2, and recipient and donor animals. CAST *Ly5* and B6 *Ly5.2* loci have the same SNPs, which are different from B6 *Ly5.1*, allowing the estimation of the level of recipient cell contamination in the donor cell populations.

Additionally, a monoclonality assay was performed by studying the expression of *Xist* to confirm that monoclonal animals were injected with a single cell. Two different SNPs in the *Xist* transcript were identified to distinguish transcription from CAST and B6 X chromosomes (**Figure 4.7 and Table 4.2**). Sanger sequencing was performed on *Xist* cDNA synthesized from extracted RNA, focusing on two strain-specific SNPs. As was foreseeable, polyclonal animals displayed two overlapping peaks in the chromatograms, whereas samples from single-HSC reconstituted animals gave rise to only one peak, corresponding either to the CAST or B6 X chromosome., suggesting that the putative monoclonal animals had indeed been reconstituted with a single cell. As for the purity assay, whole-transcriptome sequencing data was used to confirm the monoclonality of samples *in vivo* (see below). Only samples with good quality were used for further analysis.

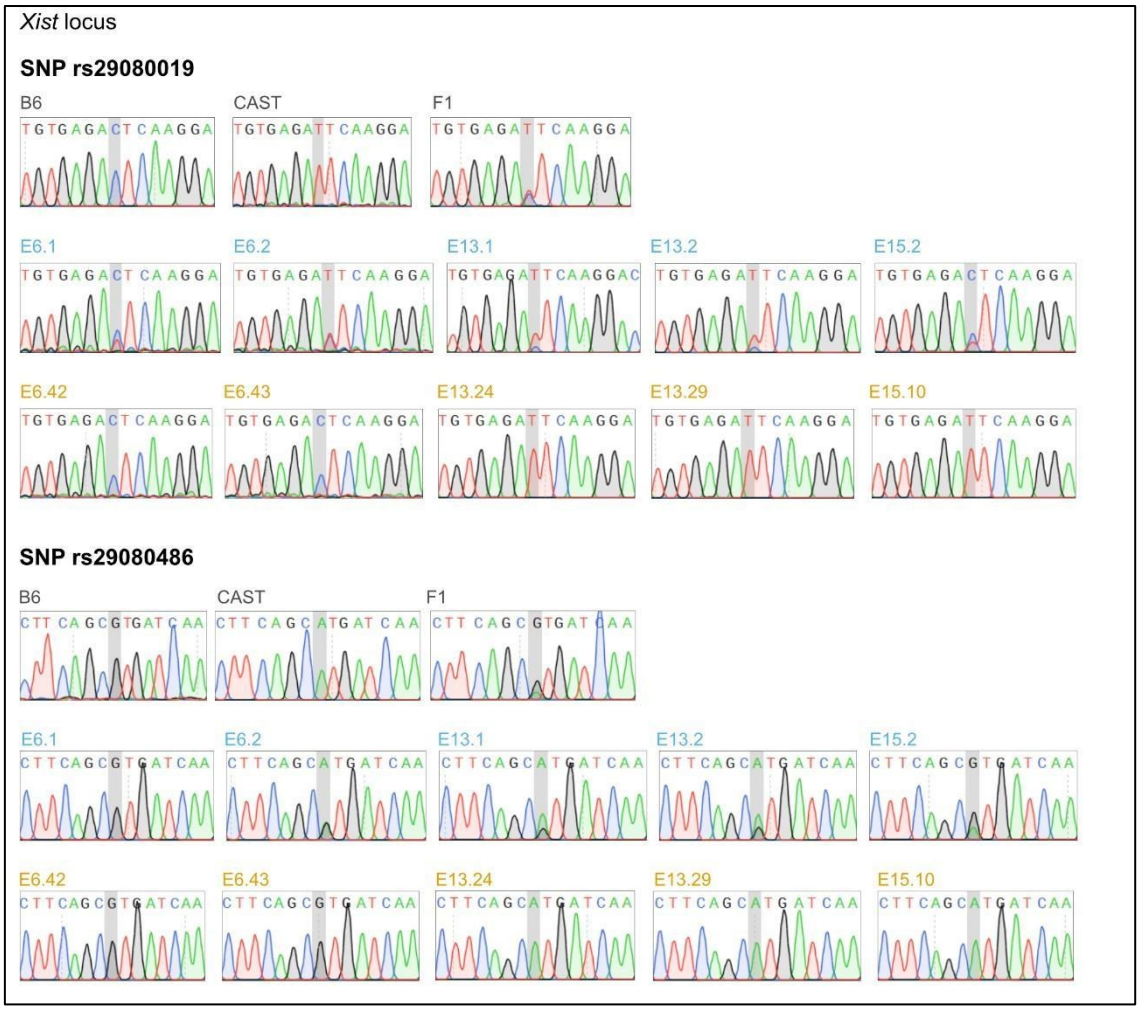


Figure 4. 7. Monoclonality assay that confirms the reconstitution of the recipient system with a single HSC. The cDNA Sanger sequencing chromatograms cover a region with two SNPs in the *Xist* locus that assigns the *Xist* transcript to the CAST or B6 X chromosome. Due to XCI, when a single cell is used for the reconstitution, a single peak is expected in the position of the SNP; when multiple cells were used for reconstitutions, two peaks should be observed in each of the SNP positions.

4.2. Transcriptome analysis

Clara F. Alves-Pereira, Alexander Gimelbrant, and Vasco M. Barreto contributed to the data analysis and graphical visualization.

4.2.1. Introduction

Whole-transcriptome sequencing or RNA sequencing (bulk RNA-seq or single-cell RNA-seq) is a popular technology used to study not only gene expression levels under different biological conditions but also novel transcripts, SNPs, insertions/deletions, alternative splicing, and splice junctions, and to understand genome function (Castel et al., 2015; Kratz and Carninci, 2014; Li, 2019). After RNA extraction, their conversion to a library of cDNA fragments and sequencing by high-throughput platforms, the obtained short sequences/reads of cDNA are mapped to a reference genome or transcriptome to calculate read counts, which are then analyzed by statistical or machine learning methods to estimate gene expression levels (Conesa et al., 2016; Li, 2019). Quality control checks should be performed at different stages of analysis to guarantee the reproducibility and reliability of results. The first quality control test starts with the analysis of produced raw reads and includes the study of sequence quality, GC content, adaptors, k-mers overrepresentation, and duplicated reads. The read alignment quality is checked by the percentage of mapped reads. This feature points to the overall sequencing accuracy and contamination with genomic DNA. Accumulation of reads at the 3' end of transcripts in samples selected for poly(A) may reveal the poor quality of starting RNA. The GC content is an indicator of PCR biases. The reproducibility among replicates should be checked for possible batch effects by using principal component analysis (PCA) (Conesa et al., 2016).

Gene expression is calculated based on read counts, i.e., the number of reads mapped to each transcript sequence. However, these raw reads alone are insufficient to compare the gene expression levels between samples. They are influenced by transcript length, the total number of reads, and sequencing biases, and need to be normalized. Normalizations such as RPKM (reads per kilobase of exon model per million reads)

(Mortazavi et al., 2008) for single-end and FPKM (fragments per kilobase of exon model per million mapped reads) (Trapnell et al., 2010) for pair-end RNA-seq remove the feature-length and library size effects, and they are suitable for within-sample comparisons. TPM (transcripts per million) (Li and Dewey, 2011; Wagner et al., 2012) is similar to RPKM and FPKM, but the order of operations is different for its calculation (first the gene length and then the sequencing depth are normalized). Due to this difference, TPM may be used for within- and between-sample comparisons for the same sample group but not for the differential expression analysis. The TMM (trimmed mean of M values) method (Robinson and Oshlack, 2010), in addition to the gene length and the sequencing depth, also takes into account biases such as heterogeneous transcript distribution; it assumes that the majority of genes are not differentially expressed and uses a weighted trimmed mean of the log expression ratios. With this normalization, it is possible to perform within- and between-sample comparisons and differential expression analysis (Conesa et al., 2016).

RNA-seq is used to study differential gene expression between two samples and differential allele expressions, also called allele-specific expression or AI. Strategies using samples highly enriched for sites with heterozygous SNPs are applied to distinguish expression levels and quantify variations between a diploid individual's two parental alleles/haplotypes. This allele-specific expression level is obtained first by retrieving allele counts from RNA-seq data over a list of heterozygous sites and then calculating the allelic ratio between reference read counts and total reads counts (Castel et al., 2015). For example, the value of 0.5 means that both maternal and paternal alleles are equally expressed, and the values of 0 or 1 correspond to expression either of only maternal or paternal allele.

RNA-seq data contain two types of variations, i.e., the technical variation introduced by protocols and machines and the biological variation caused by biological sample differences. The main challenge in RNA-seq analysis is distinguishing between real biological differences and technical noise. Different distributional assumptions to approximate the patterns of differential gene expression have been used with statistical tests to overcome this challenge. Statistical methods include parametric (Anders and

Huber, 2010; Bullard et al., 2010), nonparametric (Li and Tibshirani, 2013; Tarazona et al., 2015) and hybrid approaches (Farooqi et al., 2021). Parametric models are based on discrete distributional assumptions. First, the data are mapped to a particular distribution, and parameters are estimated for the model. After that, the future data value is predicted based on the estimated parameters. The model works properly only if the assumptions are correct, otherwise the false positive rate is highly increased. Furthermore, some of these models are influenced by outliers. This is the case when a gene is expressed in one condition but not other condition(s). Nonparametric models make fewer assumptions than parametric and use a flexible number of parameters, which increases as the model learns from more data. These models are computationally slower than the parametric models but can detect subtle data distribution features (Conesa et al., 2016; Costa-Silva et al., 2017; Farooqi et al., 2019; Huang et al., 2015). Parametric models are preferred to nonparametric models as they increase the detection power. Hybrid models, both parametric and nonparametric models, have been developed to improve the accuracy of analysis of differentially expressed genes (Farooqi et al., 2019).

Although RNA-seq becomes the tool of choice for transcriptome studies, and many software and pipelines were developed, a consensus about the best practices for RNA-seq and analysis of these types of data to obtain reproducible results has not been achieved (Koch et al., 2018; Kratz and Carninci, 2014). The first study comparing replicate runs from the same instrument of the same sample across different sequencing laboratories was performed by Li and colleagues (Li et al., 2014). It was observed that the principal source of differentially expressed genes biases was introduced by library preparation, including sample RNA isolation. It was also found that different tools for RNA-seq analysis could not remove these false positives completely. Additionally, in a more recent study where multiple replicate libraries were produced from the same RNA, varying library construction methods and the amount of sample input, it was confirmed that the principal source of overdispersion is the library preparation (Mendelevich et al., 2021). This stresses the need for appropriate experimental design and methods that can post-process and distinguish technical noise from biological differences. The majority of

studies on allele-specific expression apply the binomial test with correction for multiple hypotheses, using only one technical replicate of library preparation (Branciamore et al., 2018; Buil et al., 2015; Degner et al., 2009; Eckersley-Maslin and Spector, 2014; GTEx Consortium, 2017; Li et al., 2012; Pinter et al., 2015). However, it was demonstrated that a single replicate is insufficient to quantify the contribution of technical noise to the observed allele-specific expression unless very restrictive assumptions are taken. To overcome this challenge, a sensitive computational approach, Qllelic (https://github.com/gimelbrantlab/Qllelic), was developed to precisely calculate this technical noise by using two or more RNA-seq library replicates, decreasing the false-positive rate in AI study while conserving proper signal and increasing the accuracy and reproducibility of results. A quality correction coefficient (QCC) is calculated from comparing technical replicates. QCC reflects the concordance between replicates and is the measure of data quality. This correction effectively decreases the number of AI calls discordant between the pairs of library replicates, the number of genes identified as allelically imbalanced, and false positive rates (Mendelevich et al., 2021).

The approach mentioned before, Qllelic, was used in this study (**Figure 4.8**). RNA samples were sequenced using 2-3 library replicates, and a set of R tools was applied to obtain the AI values. Pre-processing read alignment steps to retrieve allele counts followed the ASEReadCounter* tool adapted from the GATK pipeline (Castel et al., 2015). To avoid

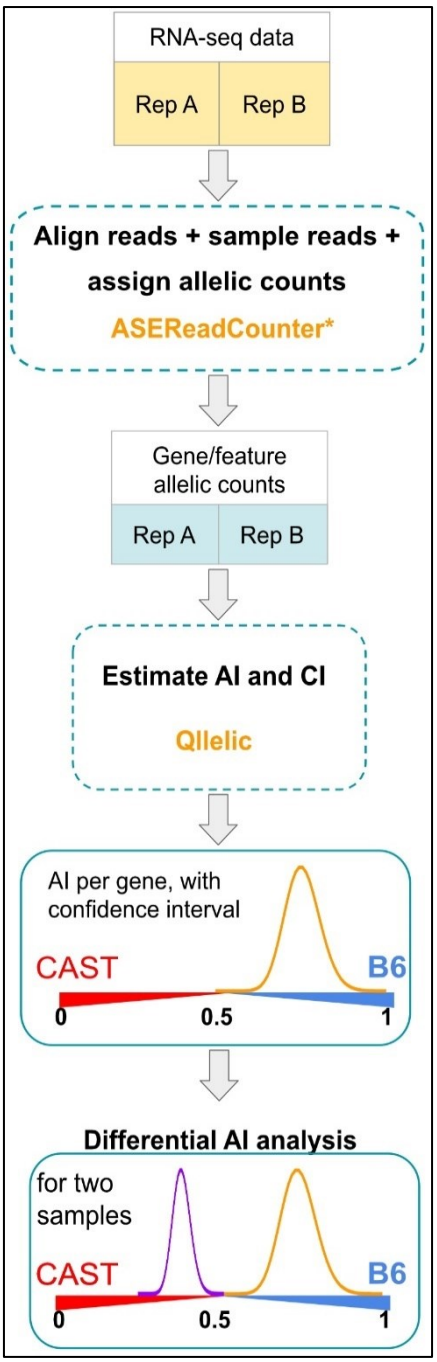


Figure 4. 8. Overview of allele-specific expression analysis. Adapted from https://github.com/gimelbrantlab/Qllelic/wiki.

bias towards the reference genome, each library replicate was aligned to two synthetic parental pseudogenomes with SNP substitutions. A read with an alternative allele contains at least one mismatch, leading to a lower probability of correct alignment than a reference read. After alignment and assignment of reads to the alleles, all replicated samples were randomly sampled to the same depth, defined by the replicate with the lowest number of reads. Finally, a table with allelic counts per gene summarizing information from SNPs separately for each replicate was computed, which then was used as input for further analysis with the Qllelic tool. QCC was calculated relying on the technical replicates and was used to correct the AI overdispersion by dividing allelic counts by QCC^2. Point estimates of AI for a gene are represented as the ratio of maternal allele counts over total allelic gene counts. The value of 1 represents the expression from the maternal-reference allele (B6 allele) and the value of 0 from paternal-alternative expression (CAST allele).

4.2.2. Quality of RNA-seq and samples

RNA samples that passed the quality test for RIN, RNA amount, monoclonality, and purity tests were 150 bp pair-end sequenced using at least two technical replicates. Raw read data were analyzed for quality with FastQC (Anders, 2010), Salmon (Patro et al., 2017), STAR (Dobin et al., 2013) and Qualimap (García-Alcalde et al., 2012) tools. The FastQC tool provides results for several features: per base sequence quality (the distribution of quality scores at each position in the read across all reads and the information whether there were any problems during sequencing), per sequence quality scores (should show that the majority of reads have high average quality scores), per base sequence content (always gives a fail for RNA-seq because of the random hexamer priming during library construction), per sequence GC content (can inform about overrepresented sequences or contamination with another organism), sequence length distribution (should present a uniform length distribution around at 150 bp, unless poor quality reads or adapter sequences were removed), sequence duplication level (relates to the complexity of the library), overrepresented sequences (can identify contamination, for example with adapter sequences), and adapter content. In general,

almost all samples passed FastQC tests, except for the test for the adapter content. However, all samples presented overrepresentation for NextEra adapters (**Figure 4.9**),

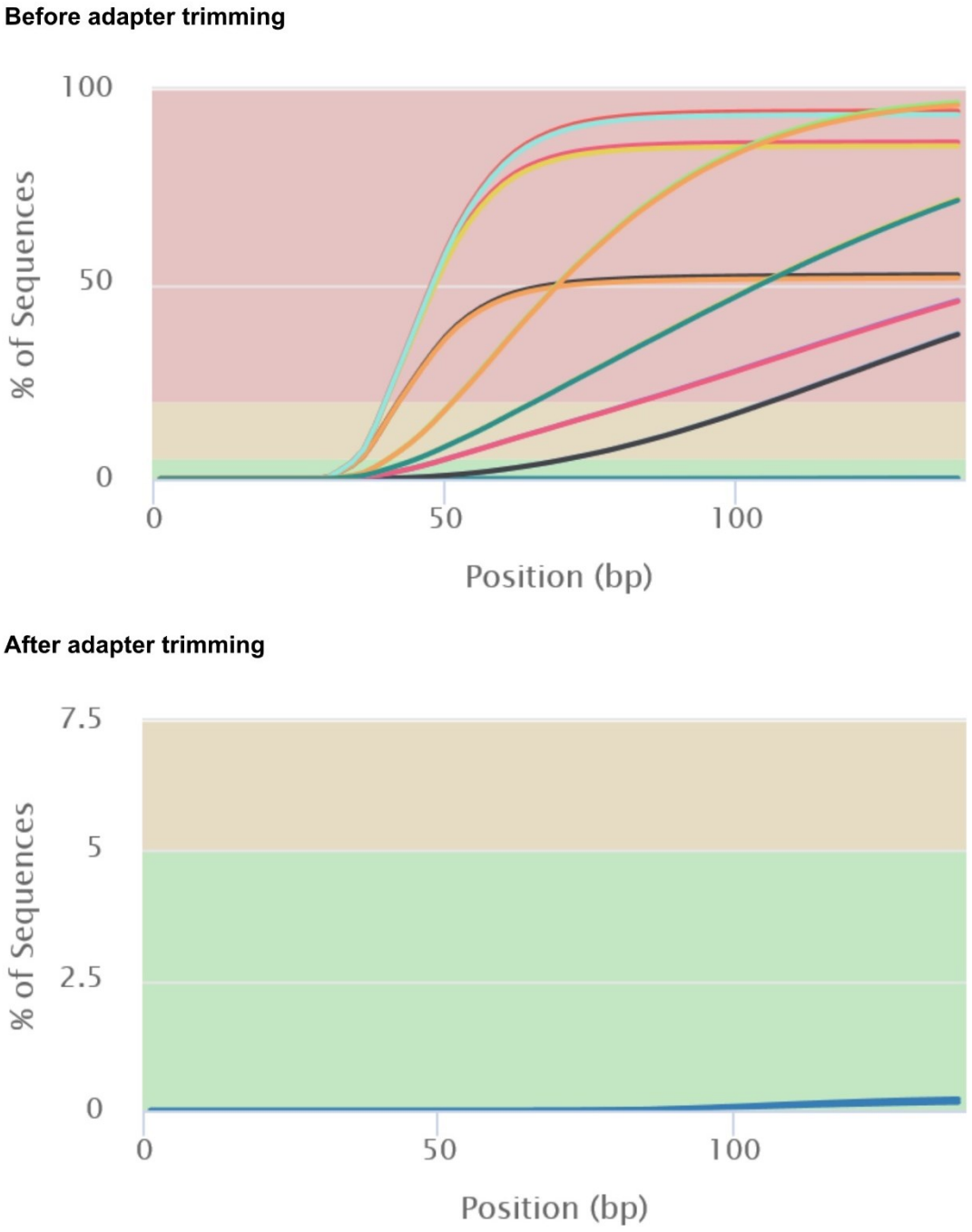


Figure 4. 9. Adapter content before and after trimming. Results were produced with the FastQC tool (Anders, 2010) and images created with MultiQC (Ewels et al., 2016) for samples control_B, control_T, E13.1_T, E13.24_B. Note that the scale of plots is different.

and even though the STAR tool (Dobin et al., 2013) used for further read alignment can account for adapter contamination and low-quality bases at the end of reads, we decided to trim these adapters, which in most cases improved the quality of the reads. Another three tools give information about the number of aligned reads to the genome, uniquely mapped reads, mapped to multiple locations, and reads aligned to exonic, intronic, and intergenic regions. These metrics were consistent across the majority of RNA samples. The identified outliers were resequenced or removed from the analysis.

The overall B cells that expanded in recipient animals reconstituted with multiple or single HSCs were sequenced from three independent experiments (E6, E13, and E15). Each of them had different donor animals, which were used for reconstitutions. From experiment 13, T cells were also sequenced. As non-clonal controls, B and T cell populations from an unmanipulated donor animal were used for analysis. In total, seven polyclonal, seven monoclonal, and two non-clonal samples contributed to the final RNA-seq study (**Figure 4.10**).

As was mentioned above, we deepened the analysis of monoclonality using RNA-seq data to calculate the allele-specific expression of the X-linked genes (**Figure 4.10**, see also **Figure 4.22 A**) and contamination of sorted donor cell populations with recipient cells (**Figure 4.11**). Allele-specific expression is represented by AI values, which are calculated as a ratio between the maternal allele reads and total allele reads (i.e., AI = maternal allele reads / (maternal + paternal allele reads)). Thus, as mentioned before, AI values vary between 0 and 1, with AI=1 for absolute expression from the maternal allele, AI=0 for paternal allele, and AI=0.5 for equivalent expression from both alleles. As expected, the median AI values of the X-linked genes for unmanipulated and polyclonal samples are balanced (0.5±0.2). In contrast, the median AI values obtained from animals reconstituted with a single HSC are extreme. The animals that inactivated the maternal X chromosome (B6), namely E6.42 and E6.43, present AI values slightly above zero (0.02±0.01), and mice that inactivated the paternal X chromosome (CAST), E13.24, E13.29, and E15.10, have AI values close to one (0.96±0.03). However, the AI values of B cells from E13.24 and E13.29 animals seem to deviate from the expected value of one and express the paternal X chromosome in low amounts. The scenario in which two or

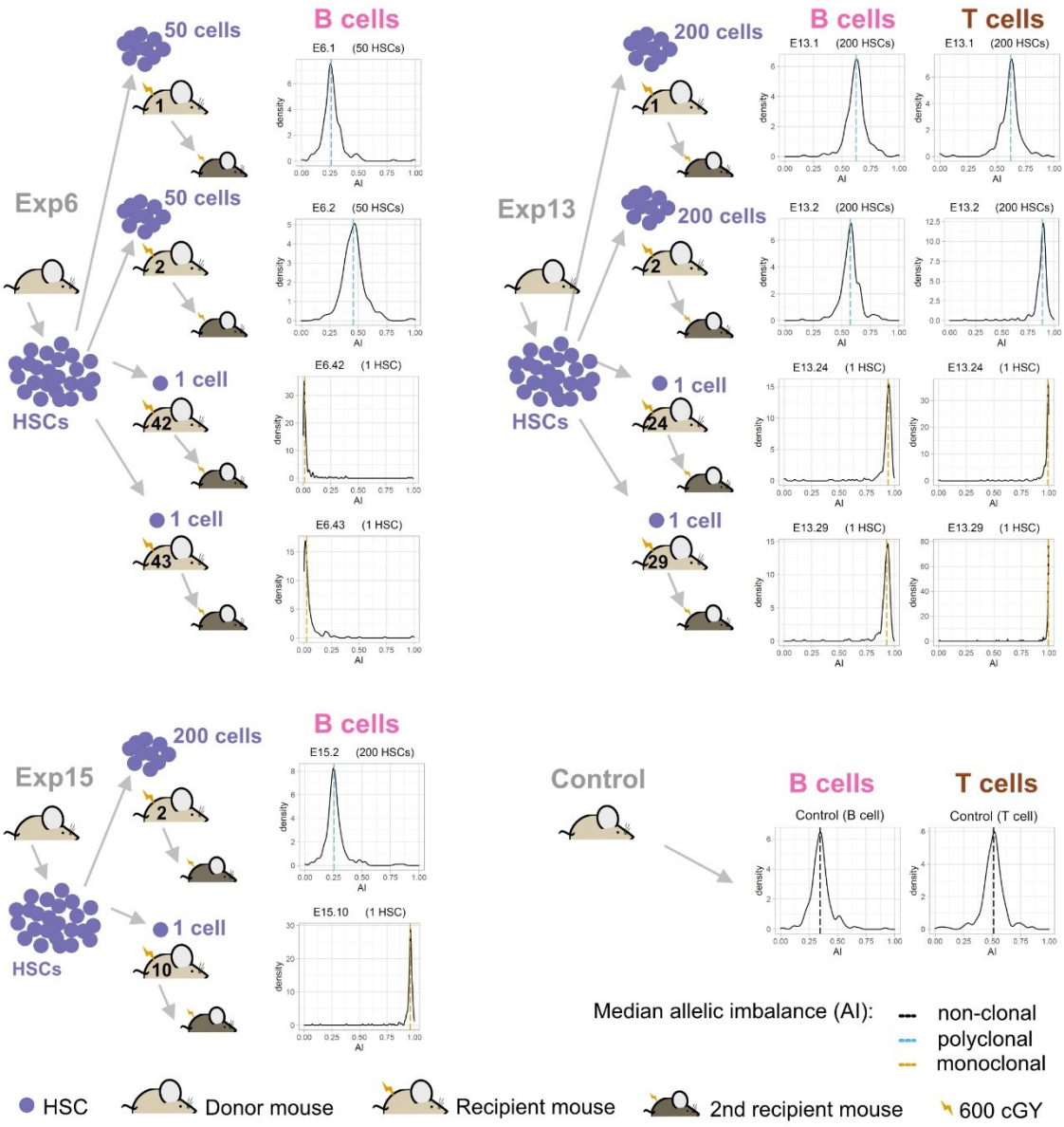


Figure 4. 10. Overview of single and multiple HSC reconstitutions that originated the samples used for RNA-sequencing (experiments E6, E13, and E15). In each experiment, HSC cells isolated from one donor mouse F1($CAST^{Ly5/Ly5}$ x $B6^{Ly5.2/Ly5.2}$) were injected in multiple sub-lethally irradiated recipient animals F1($CAST^{Ly5/Ly5}$ x $B6^{Ly5.2/Ly5.1}$). Different donors were used for each experiment. All animals showed long-term reconstitutions, and both monoclonal and polyclonal cells from primary repopulated animals reconstituted a secondary recipient. The density plots represent the allelic ratios of X chromosome-linked genes for each sample, as measured by RNA-Seq.

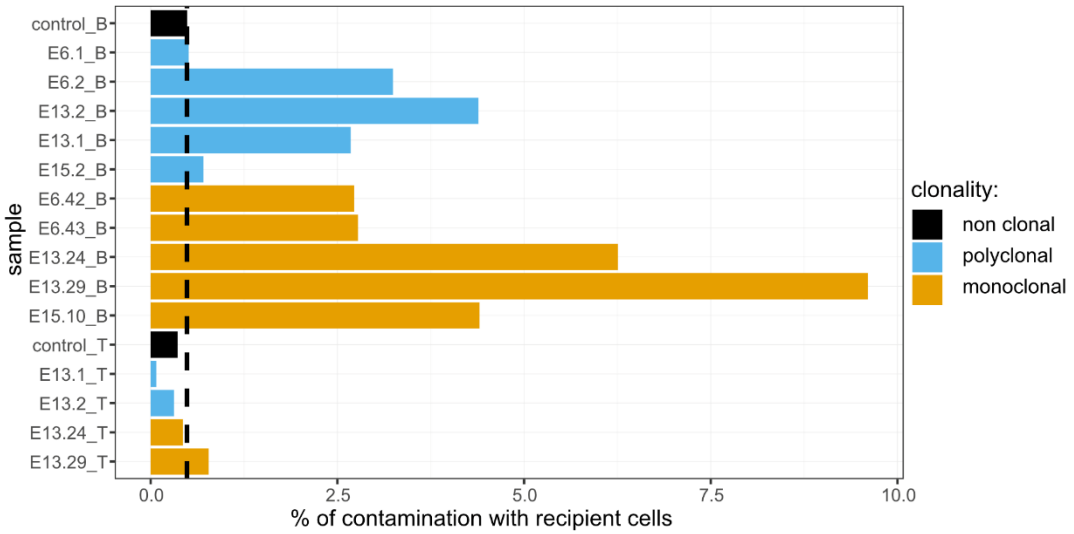


Figure 4. 11. Estimation of donor population contamination with recipient cells using RNA-seq. Percentages obtained from next-generation sequencing of recipient cells in the sorted donor cell populations focusing on three different SNPs for the Ly5 gene that distinguish two pan-leukocytic markers, Ly5.1 and Ly5.2, that allow us to identify recipient and donor cells, respectively. The nucleotide bases for Ly5.1 and Ly5.2 were counted for each SNP, considering that CAST and Ly5.2 have the same SNPs, and the average percentage of Ly5.1-recipient cell contamination was calculated. The dashed line (0.5%) represents the percentage of artifactual SNPs due to errors introduced by sequencing, which was estimated by the sequencing results of the unmanipulated donor mouse.

more HSCs were injected in these recipient animals is excluded because T cell samples from the same animals do not have this leakage (the AI value for E13.24_T is 0.99 and for E13.29_T is 1.00). The most likely explanation is the contamination of the sorted samples with recipient Ly5.1 cells, a polyclonal population of cells. To test this hypothesis, Ly5.1 and Ly5.2 SNPs were quantified in the RNA-seq data (**Figure 4.11**). By calculating the presence of Ly5.1 recipient SNPs in the unmanipulated control donor animals, it was estimated that sequencing errors only contributed to around 0.5% of SNPs. Six samples have around 1% of contaminating recipient cells (E6.1_B, E13.1_T, E13.2_T, E13.24_T, E13.29_T, E15.2_B), six samples have between 2.5% and 5% (E6.2_B, E6.42_B, E6.43_B, E13.1_B, E13.2_B, E15.10_B), and the E13.24_B and E13.29_B samples have the highest percentages of contaminating cells, i.e., 6.3% and 9.6%, respectively. This means that samples have low and, in two cases, relatively high

recipient cell contamination, but in any case, this contamination, as well as possible genomic DNA contamination, will never overestimate cases of allelic-specific expression and lead to erroneous conclusions. Instead, if anything, it will lead to a slight underestimation of RMAE. Since the contamination of the samples correlates with the deviation from the expected AI values in the monoclonal animals (**Figure 4. 11**), we can conclude that the monoclonal samples were expanded *in vivo* from a single HSC. This conclusion is also supported by the observation that the frequencies of single-HSC reconstitution are low (**Table 4.1**), which makes extremely low the probability of having two HSCs introduced by mistake in the same recipient.

Additionally, to check the expansion dynamics of produced lymphoid lineage samples from HSC, a study of antigen receptor V(D)J rearrangement clonotypes was performed using RNA-seq data (**Figure 4.12**). Similar rearrangements are observed in samples expanded from single and multiple HSC reconstitutions, and also unmanipulated control samples. This means that all samples are equivalently complex in terms of the V(D)J repertoire. Furthermore, there was an abundant cellular proliferation of single-HSC derived lymphoid cell populations before V(D)J rearrangement, which starts in pro-B and pro-T cells.

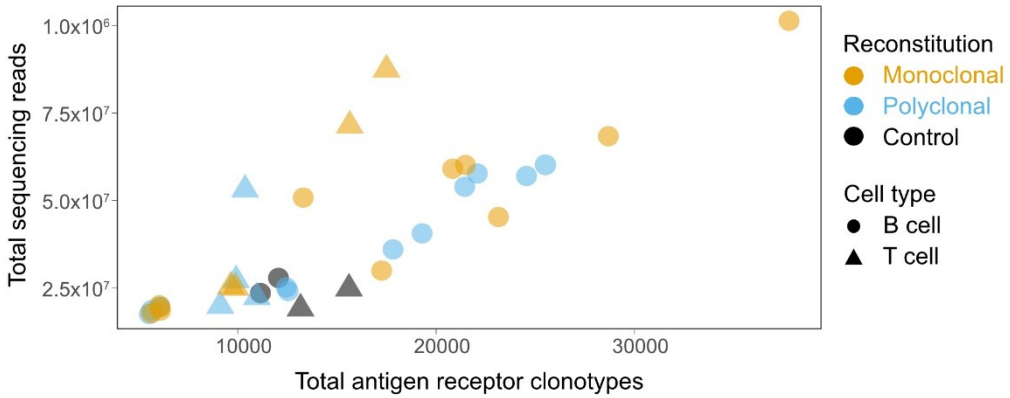


Figure 4. 12. The complexity of the VDJ repertoire in sequenced B and T samples. VDJ clonotypes in different populations of donor B and T cells expanded *in vivo* and the control animal. The number of VDJ rearrangements identified with the MiXCR tool (Bolotin et al., 2015, 2017) on each sample (x-axis) were plotted against the number of sequenced reads (y-axis).

4.2.3. Identification of autosomal allele-specific expression

After reading alignment and estimation of allelic expression, a total of 27,510 genes were detected. From this list, we removed genes with substantially low expression (<10 TMM-normalized counts), loss of heterozygosity (LOH), obtained from whole-exome sequencing data, and genetic biases, which can be due to different mouse strain genetics or parental imprinting (i.e., genes with extreme expression from the same allele in all samples: monoclonal, polyclonal and control samples). Finally, 8,441 genes in B cells and 8,377 in T cells were detected; 7,486 genes are common to both tissues, 955 genes are B cell-specific, and 891 are T cell-specific (**Figure 4.13**).

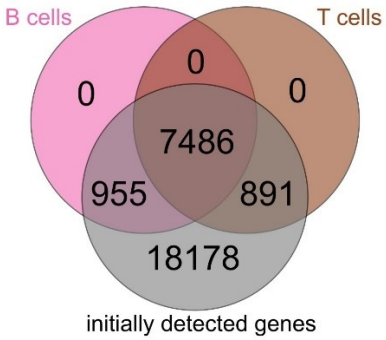


Figure 4. 13. Venn diagram representing the overlap between the initially identified genes and genes used in the allele-specific expression analysis. Genes with no expression, loss of heterozygosity, and genetic biases were removed to avoid an overestimation of allelic imbalance.

Instead of simply classifying genes into discrete categories such as balanced or imbalanced, a differential AI analysis was performed to address a more quantitative and interesting question. This analysis estimates how a gene is expressed differently from a specific allele between two samples. It is expected that the AI in polyclonal samples is more balanced and that the AI in monoclonal samples may deviate from the AI in polyclonal samples, particularly in the case of RMAE. Thus, a pairwise AI comparison was produced between all samples: polyclonal against monoclonal, polyclonal against polyclonal, and monoclonal against monoclonal (**Figure 4.14 A** and **Supplementary Figure 7.1**). Samples with equal AI values should align all genes over the diagonal and produce a Pearson's coefficient correlation (r^2) close to 1. In contrast, samples with

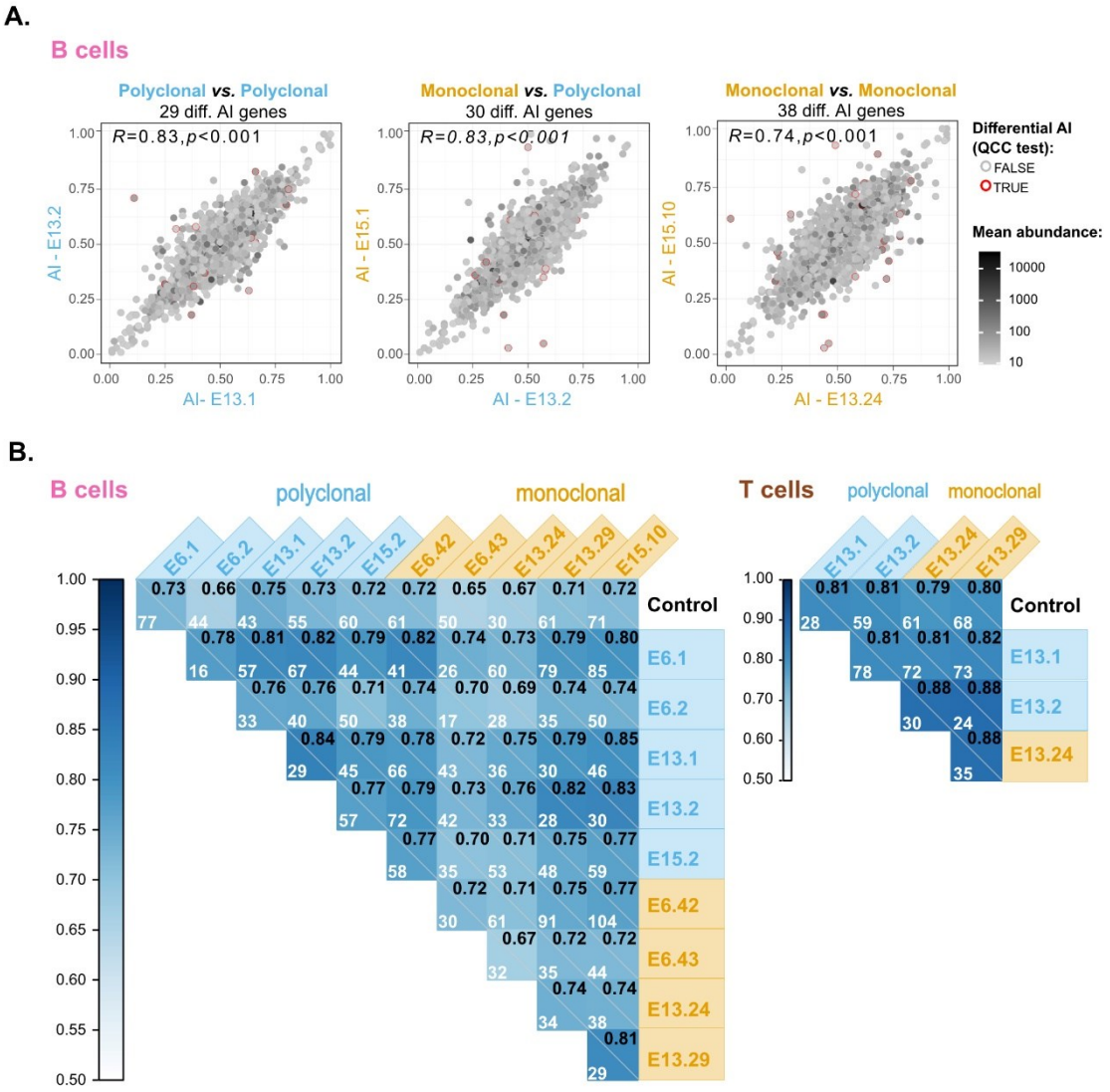


Figure 4. 14. Comparison analysis of samples to search for genes that maintain allelic imbalance during hematopoietic differentiation. (A) Representative dot plots of pairwise comparisons of AI between monoclonal vs. polyclonal samples, polyclonal vs. polyclonal samples, and monoclonal vs. monoclonal samples. The red circles signal the genes for which differential AI remained statistically significant after QCC correction on the binomial test. The total number of these genes per comparison is shown above each plot. The Pearson's coefficient correlation for all AI pairwise comparisons is also shown at each dot plot's upper left corner. A greyscale coloring the dots represents the mean expression between the two samples, calculated from each sample's TMM-normalized counts. (B) Correlograms for B and T samples. Pearson's coefficient correlation of AI for all pairwise comparisons between samples. Pearson's coefficient is represented in the upper right corner within each square, and the number of genes with a significant differential AI in each pairwise comparison after applying QCC correction on the binomial test is also shown.

differential AI should present more genes that deviate from the diagonal and produce lower r^2 values. To test this hypothesis, a correlogram with r^2 and the number of genes with significant differential AI in each comparison for B and T samples after applying QCC correction on the binomial test was created (**Figure 4.14 B**). It is observed that comparisons involving at least one monoclonal sample are similar to other comparisons, meaning that most autosomal genes with AI present in hematopoietic stem states are not maintained after extensive abundant cellular expansion and differentiation.

To confirm that monoclonal samples are identical with polyclonal or control samples, a t-distributed stochastic neighbor embedding (t-SNE) analysis, which is used to visualize high-dimensional data in a low-dimensional space (Van Der Maaten and Hinton, 2008), was conducted (**Figure 4.15**). In terms of AI in autosomal genes, this analysis did not cluster polyclonal and control samples, which should be similar. Furthermore, it did not display the expected scattered distribution of monoclonal samples if each clonal cell line kept distinct allele-specific expression. Thus, we conclude that all samples are very similar and if some differences are present, t-SNE analysis is not able to reveal them.

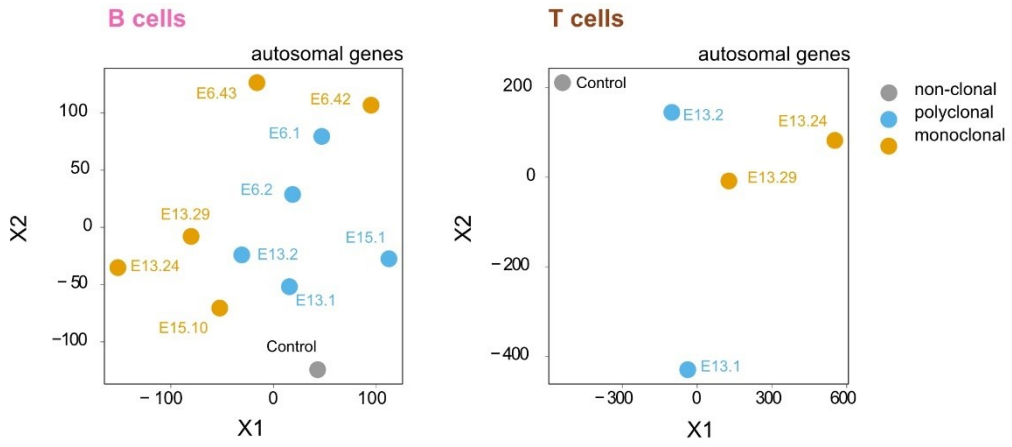


Figure 4. 15. Visualization of high-dimensional data of autosomal allelic imbalance in a low-dimensional space using the t-SNE algorithm to compare the dispersion of polyclonal and monoclonal samples.

An alternative analysis to dissect differences between monoclonal and polyclonal samples is an exploration of AI dispersion (e.g., standard deviation). It is expected that, in the group of polyclonal samples, any given gene will tend to show balanced allele-

specific expression and therefore a low dispersion of AI, whereas the AI of a gene under RMAE in the group of monoclonal samples will have a higher dispersion. The B cell samples, which are more abundant than the T samples, were used to calculate the ratio of standard deviation from the polyclonal and monoclonal groups, expressing it as a ratio of standard deviations. By performing a one-sided Wilcoxon test, it is possible to conclude that the standard deviation of AI in the monoclonal group is greater than in the polyclonal group for a small set of genes ($p < 2.7 \times 10^{-6}$; **Figure 4.16 A**). X-linked genes were also plotted as a control to confirm that this plot visualization can identify genes with variable AI. The 14 genes that show the highest monoclonal and polyclonal AI standard deviation ratio were highlighted (arbitrarily set at an AI SD of 0.15). To investigate the behavior of these 14 genes over all samples (**Figure 4.16 B**), the individual AI values of all polyclonal and monoclonal samples were plotted over those of the control unmanipulated animal (green circles), and it is clear that the dispersion of AI values in monoclonal samples is higher than in polyclonal samples. The *Pkp3* transcript is an interesting and the most solid example of this high dispersion. In some monoclonal samples, this transcript is predominantly expressed from maternal alleles, in others from the paternal allele, and in one case it also has balanced expression from both alleles. This means that *Pkp3* carries random allele-specific expression preserved over extensive differentiation steps already present in the original HSC. B and T cells share the same lineage way of differentiation until the CLP stage and then split into two independent bifurcations. Thus, if these allele-specific states present in the identified 14 genes were already established during the hematopoietic stem state and then remained stable during lineage commitment, the AI values for T and B cells from the same animal should be positively correlated. A pairwise comparison between the same animal but different cell types was performed (i.e., E13.24_B against E13.24_T and E13.29_B against E13.29_T). It is clear that the AI values from different cell types derived from the same HSC are similar, as most genes are plotted along the diagonal ($R=0.85$, $p<0.001$). (**Figure 4.16 C**). Pairwise control analysis of AI values from different cell types derived from different HSCs was performed and shows that allele-specific states are very different ($R=0.33$, $p=0.147$). This is consistent with the above notion of clonally stable allele-specific states.

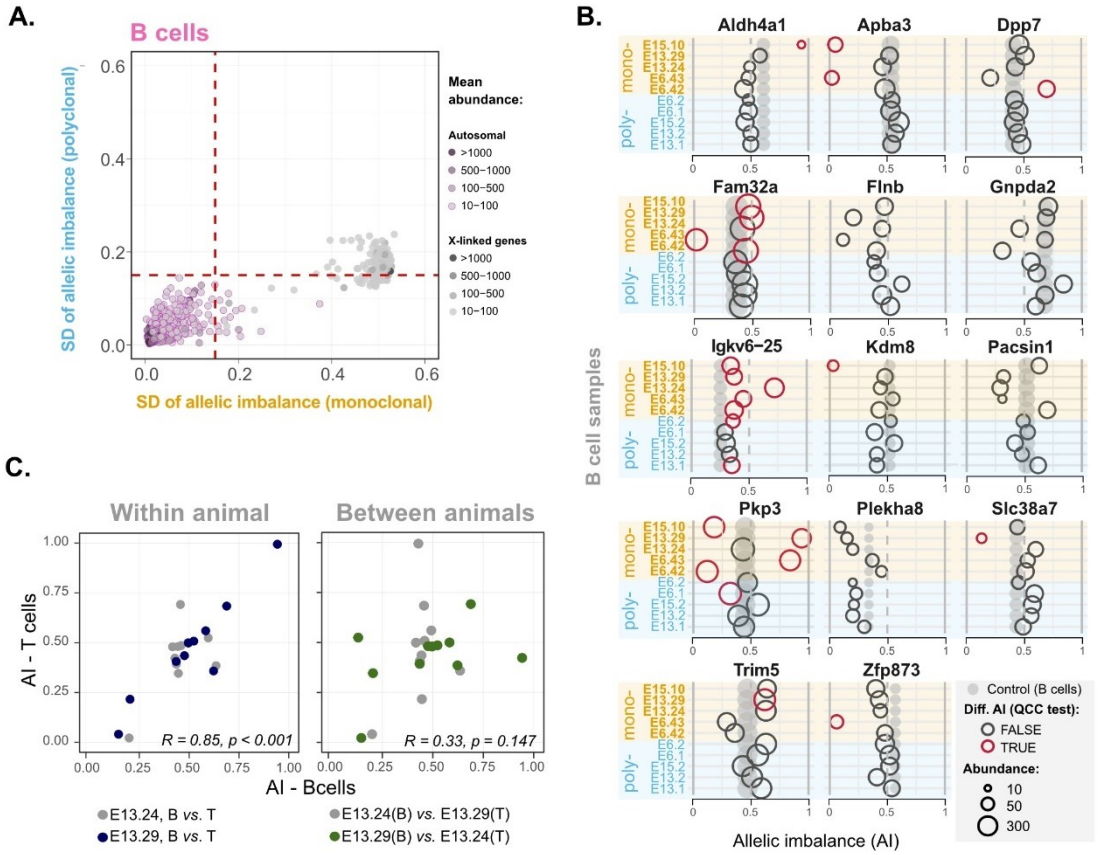


Figure 4. 16. Allele-specific states for some genes are stable and persistent over extensive cell expansion and differentiation from the hematopoietic stem state. (A) Dot plot showing standard deviations (SD) of AIs for five B-cell monoclonal samples (x-axis) against the SD of AIs for five polyclonal samples (y-axis). Dashed vertical and horizontal lines - arbitrarily set at an AI SD of 0.15 - represent the threshold above which genes were considered potentially intrinsically imbalanced. Pink-circled dots represent the autosomal genes, and uncircled dots represent the X-linked genes (control). Only genes for which differential AI remained statistically significant after QCC correction in at least one pairwise comparison (i.e., the red dots in Figure 14) within monoclonal B samples or polyclonal B samples and with expression in all B-cell samples are shown. Abundance values are TMM-normalized counts. (B) Comparison of putative transcriptionally stable allelically imbalanced genes between all samples and non-clonal control B. Grey dots are AIs of the unmanipulated animal control sample, and empty circles are AIs of monoclonal or polyclonal samples. Red circles represent comparisons for which AI differences remained statistically significant after QCC correction for control B comparison. The diameter of dots/circles is proportional to the abundance (in TMM-normalized counts). (C) Dot plots show the AI of putative transcriptionally stable allelically imbalanced genes in B cells (x-axis) against those in T cells (y-axis). Pairwise comparisons for two monoclonal animals are shown. In the left plot, each animal's B and T cell data are paired (within animal comparison). In contrast, the right plot is an artificial control in which the B and T cell data from different animals are paired (comparison between animals). Each plot shows the Pearson's coefficient correlation considering the combined animal datasets; the Pearson's coefficient correlations for each animal dataset are R=0.33 (p=0.147) and R=0.85 (p<0.001).

Another explanation for these stable allele-specific variations is that, instead of RMAE, these genes underwent a LOH genetic modification such as a deletion, already carried by the original HSC and then perpetuated throughout differentiation. To confirm that the identified genes with AI present both CAST and B6 exons, exome analysis was carried out for three animals for which exome-sequencing data were produced. The AI calculated from RNA-seq data was plotted against AI from exome-seq data for each animal, and it is observed that AI from DNA data is balanced and, once more, AI values from RNA data of monoclonal animals (E6.43 and E15.10) show high dispersion than the polyclonal animal (E6.2). This means that the CAST and B6 sets of exons are present in the genome. Thus, the observed AI variation at the transcript level is unlikely to result from genomic deletions (**Figure 4.17**).

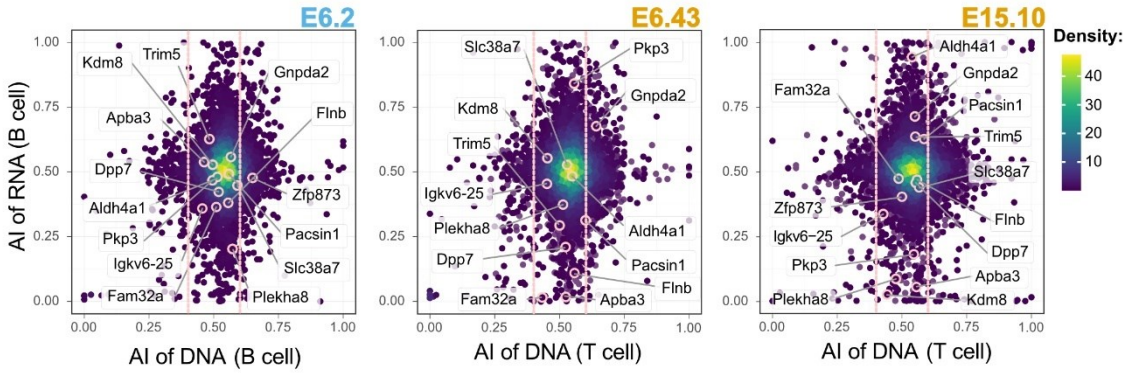


Figure 4. 17. Loss of heterozygosity analysis of putative transcriptionally stable allelically imbalanced genes. AI from RNA-seq data plotted against AI from whole-exome sequencing data for the same animals (polyclonal sample E6.2, and monoclonal samples E6.43 and E15.10). Only genes with abundance>10 TMM-normalized counts are represented. For the DNA axis (x-axis), all of these genes fall in the vicinity of the dotted vertical lines highlighting the 0.4–0.6 AI balanced range.

Given the importance of excluding LOH as an explanation for the rare cases of RMAE, we performed a bootstrapping analysis (100,000 replicates per distribution) to complement the previous exome sequencing analysis. Essentially, we evaluated the likelihood of randomly finding a group of genes with the mean difference between the AIs in DNA and RNA data (AI_{DNA} - AI_{RNA}) as high as the mean value for the rare 14 genes (**Figure 4.17**) in the set of monoclonal animals (**Figure 4.18**). If the highly variable AI values of RNA have epigenetic bases and are not the result of LOH, the mean difference

between AI values obtained from transcriptomics and exome sequencing data should be notably high. In contrast, if the somatic rearrangement events leading to LOH are the explanation for the RMAE patterns we observed, at the level of DNA the data will mimic the transcriptomics data and, therefore, the AI_{DNA} - AI_{RNA} difference will be low. The bootstrapping revealed that, for E6.43 and E15.10, random sampling is unlikely to produce a group of genes with higher AIDNA - AIRNA mean differences than the ones we found (p=0.0003 and p=0.0002, respectively).

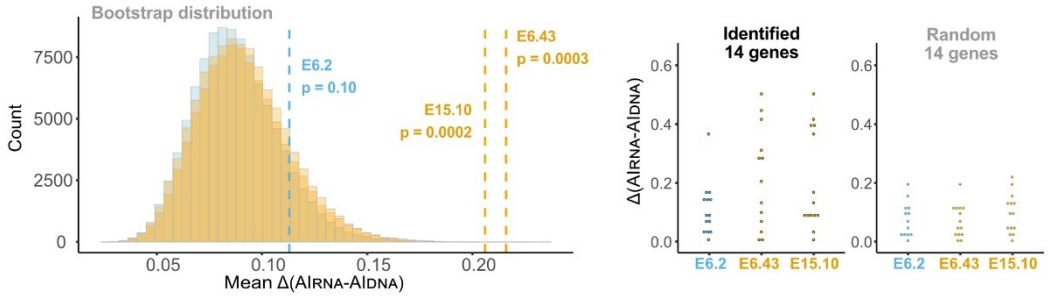


Figure 4. 18. Bootstrapping analysis of difference between the AIs in DNA data and RNA data (AI_{DNA} - AI_{RNA}) in two monoclonal samples for the genes with persistent clone- and allele-specific autosomal transcriptional states (highlighted in 4.16 B). In the left panel, the histogram represents the distributions of the means of the difference for 13 or 14 randomly sampled genes generated by bootstrapping the transcriptomics data (100,000 replicates per distribution). The dashed lines show the observed AI_{DNA} - AI_{RNA} means for the 13 and 14 of the 14 putative transcriptionally stable allelically imbalanced genes detected in the monoclonal samples E6.43 and E15.10, respectively, which are statistically different from the mean of a random sample considering the respective distributions (p=0.0003 and p=0.0002, respectively), unlike the AI_{DNA} - AI_{RNA} mean for the 14 putative transcriptionally stable allelically imbalanced genes in the E6.2 polyclonal sample (p=0.10). The right panel shows the distribution of the | AI_{DNA} - AI_{RNA} | observed for the putative transcriptionally stable allelically imbalanced genes and a random sample of size 14 in E6.2, and E15.10, and 13 in E6.43.

Genetic variations are also associated with replication fragile sites. It is known that these regions have common molecular features such as high expression levels and large size (Barlow et al., 2013; Helmrich et al., 2006). Our analysis of the *locus* size, open reading frame (ORF) size, and expression levels in LT-HSCs (publicly available data) suggests that the genes with persistent clone- and allele-specific autosomal transcriptional states across), are not associated with the features of replication fragile sites (**Figure 4.19**).

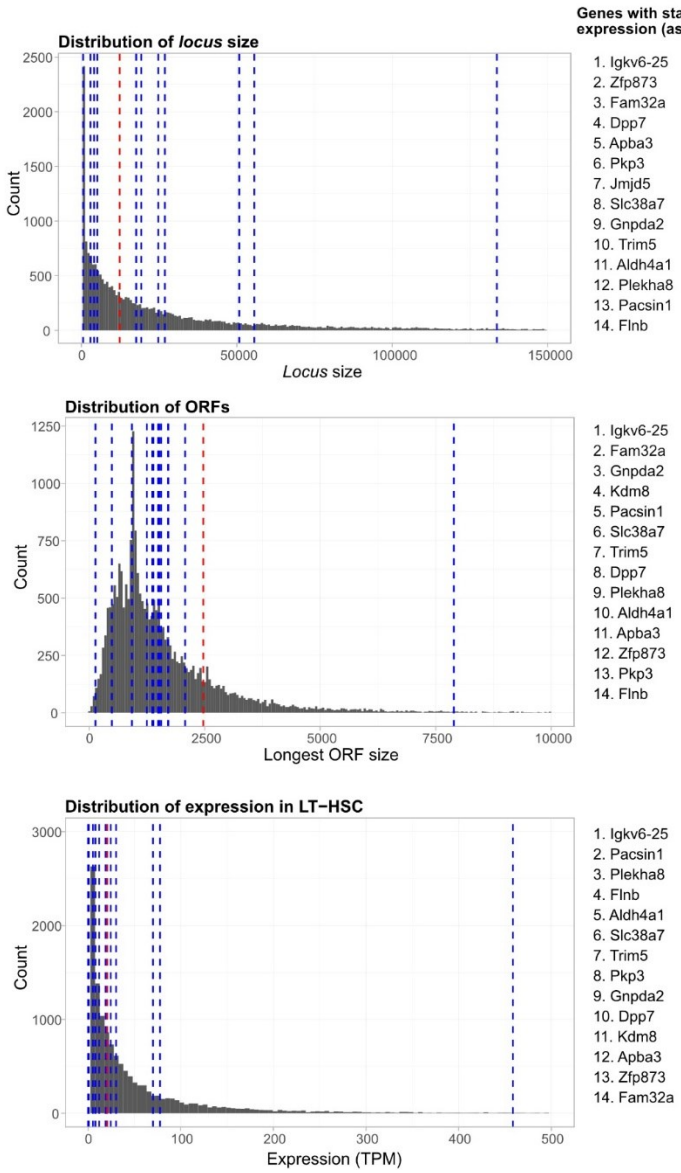


Figure 4. 19. Association of genes with persistent clone- and allele-specific autosomal transcriptional states with common molecular features related to replication fragile sites. Location of 14 genes across distributions of *locus* size of all protein-coding genes, open reading frame (ORF) size, and expression in LT-HSCs. Gene sizes were obtained from the gencode mouse genome downloaded GTF file (http://ftp.ebi.ac.uk/pub/databases/gencode/Gencode_mouse/release_M27/gencode.vM27.annotation.gtf.gz) with custom scripts. ORFs were generated from the downloaded genecode transcript sequences fasta file (https://ftp.ebi.ac.uk/pub/databases/gencode/Gencode_mouse/release_M27/gencode.vM27.transcripts.fa.gz) using the orfipy tool (Singh and Wurtele, 2021) with the standard codon table and default parameters. The longest ORF for each gene was plotted for distribution. Expression in LT-HSC was obtained from the Immunological Genome Project (https://www.immgen.org/), GEO:GSE109125. *Locus* and expression plots were zoomed-in. The blue lines correspond to genes with stable allele-specific transcription through HSC differentiation and the red line corresponds to the gene *Pkp3*.

We then compared the frequencies of genes under RMAE in clones that underwent extensive differentiation (this study) and clones that underwent cellular proliferation without extensive differentiation (**Table 1.2**) (Branciamore et al., 2018; Eckersley-Maslin et al., 2014; Gendrel et al., 2014; Gimelbrant et al., 2007; Jeffries et al., 2012, 2016; Li et al., 2012; Pinter et al., 2015; Zwemer et al., 2012). This discrepancy could be due to the different experimental and statistical approaches used in this study. To eliminate this source of variation, the same RNA-seq pipeline analysis was applied in data produced from clonal cells without undergoing differentiation. v-Abl pro-B clonal cell lines Abl.1, Abl.2, Abl.3, and Abl.4 were derived previously from 129S1/SvImJ x CAST/EiJ F1 female mice by expansion of FACS-sorted single cells after immortalization (Zwemer et al., 2012). These clones were sequenced using two technical replicates of the cDNA library per sample (Gupta et al., 2021) and analyzed using the same QCC correction on a binomial test to exclude false positives (Mendelevich et al., 2021). LOH obtained from whole-exome sequencing (Gupta et al., 2021) was excluded as well as genes with no expression (<10 TMM-normalized counts) and genetic biases. Pairwise analysis of differential AI values was performed the same way as for HSC-derived clones (**Figure 4.20 A, B** and **Supplementary Figure 7.2**), and it was found that the clones that expanded without differentiation show at least four times more genes with significant differential AI in each comparison than HSC-derived clones with extensive differentiation (**Figure 4.14**). Additionally, to perform a similar AI dispersion analysis, the AI standard deviation of four Abelson clones was plotted against the AI standard deviation of four polyclonal HSC-derived cells (since the standard deviation depends on sample size, the comparison has to be four against four (**Figure 4.20 C**)). It is clear that the Abelson clones show more genes with variable AI than the monoclonal HSC-derived clones. These results suggest that clones undergoing extensive differentiation lose allele-specific states, unlike the clones that are kept *in vitro* without differentiation.

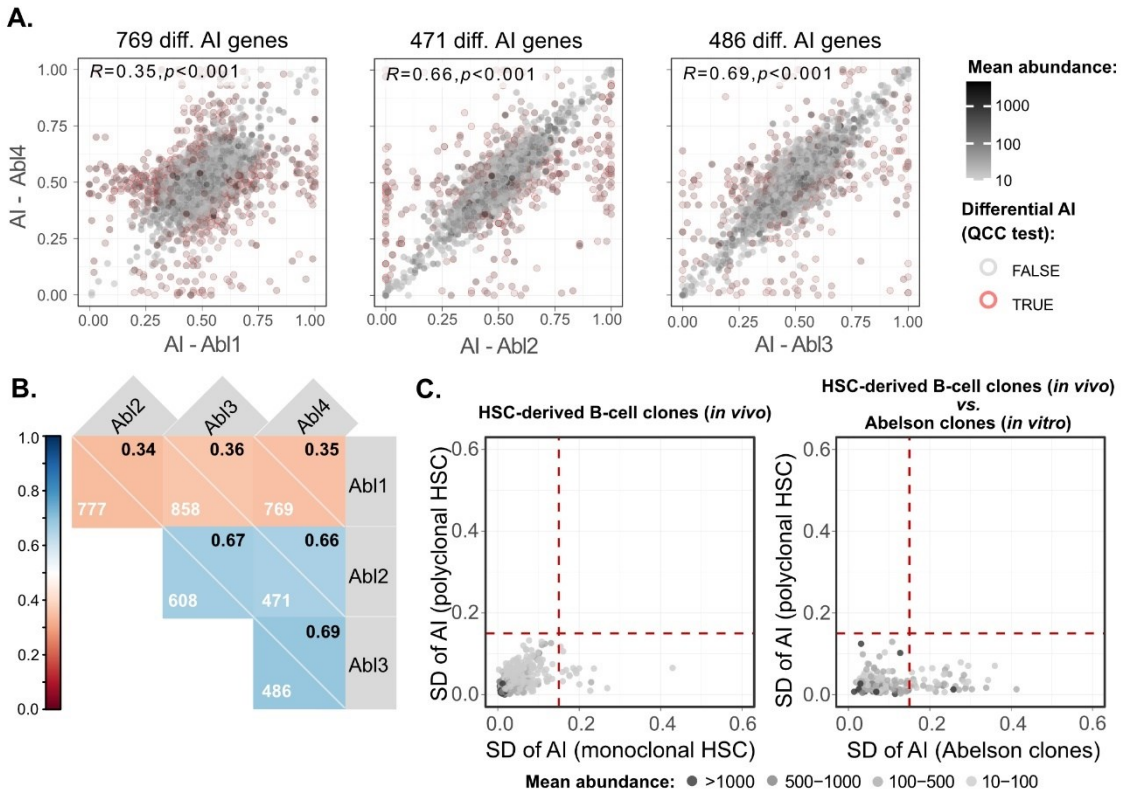


Figure 4. 20. B clones expanded *in vitro* show more genes with clonal-specific AI than B cells differentiated from a single HSC *in vivo*. (A) Representative dot plots of pairwise comparison of AI between different Abelson-immortalized B-cell clones. Red circles signal the genes for which differential AI remained statistically significant after QCC correction on the binomial test. The total number of these genes per comparison is shown above each plot. The Pearson's coefficient correlation for all AI pairwise comparisons is also shown at each dot plot's upper left corner. Mean abundance levels (mean TMM-normalized counts) are continuous greyscale colors. (B) Correlogram with pairwise comparisons of Abelson-immortalized B-cell clones. Pearson's coefficient correlation of AI for all pairwise comparisons between samples. Pearson's coefficient is represented in the upper right corner within each square, and the number of genes with a significant differential AI in each pairwise comparison after applying QCC correction on the binomial test is also shown. (C) Two dot plots showing standard deviations (SD) of AIs for four monoclonal (x-axis) against four polyclonal (y-axis) HSC-derived B cell samples (left plot), and SD of AI for all four Abelson clones (x-axis) against the SD of AI for four polyclonal HSC-derived B cell samples (y-axis) (right plot). Whole-exome sequencing data were used to exclude transcripts with possible LOH. Dashed vertical and horizontal lines set arbitrarily at an AI SD of 0.15 represent the threshold above which genes were considered potentially intrinsically imbalanced. Mean abundance levels (mean TMM-normalized counts) are represented as binned greyscale colors.

4.2.4. Identification of XCI escapees

Finally, the strategy used in this work to identify genes under RMAE could be suitable to study genes that are expressed from both active and inactive X chromosomes, which are known as XCI escapees. A correction with QCC, calculated from two or more technical replicates, was applied to the RNA-seq data from clonal samples produced *in vivo* to identify murine hematopoietic lineage-specific XCI escapees. For this strategy (**Figure 4.21**), we converted AI values of X-linked genes relatively to inactive X chromosome (i.e., AI = inactive X chromosome-linked allele reads / (inactive + active allele reads), corresponding to 0 as no expression and 1 to the expression of the inactive X chromosome). We then identified genes with expression from the inactive X chromosome of at least 10% of total expression (Carrel and Willard, 2005). For instance, sample E13.24_T has an AI value of 0.99 for the X-linked genes, which means that the maternal X chromosome is active. To convert this value to search for the genes that have expression from inactive X chromosome above 10% of the total expression level, the expression of each gene from this sample was compared with a sample-corrected threshold of 0.11 calculated as: 0.01 (i.e., 1 - 0.99 to obtain the AI median value of X-linked genes from the inactive X chromosome in this sample and include a fraction of AI corresponding to sequencing errors, and recipient cell contamination) + 0.1 (i.e., 10% of the inactive X chromosome). For samples with an active paternal X chromosome, the conversion of AI values for the inactive X chromosome is unnecessary, as values of 0 correspond to the inactive X chromosome. After that, the binomial test and QCC correction were applied to classify a gene as significantly different from this threshold in the sample. A gene was categorized as XCI escapee if it met three criteria: 1) only samples with abundance higher than 10 TMM-normalized counts were considered; 2) the median of AI in the control samples (polyclonal and control samples) was balanced (0.5±0.2); and 3) the AI of the gene is statistically different from the threshold in at least two samples, irrespective of the tissue.

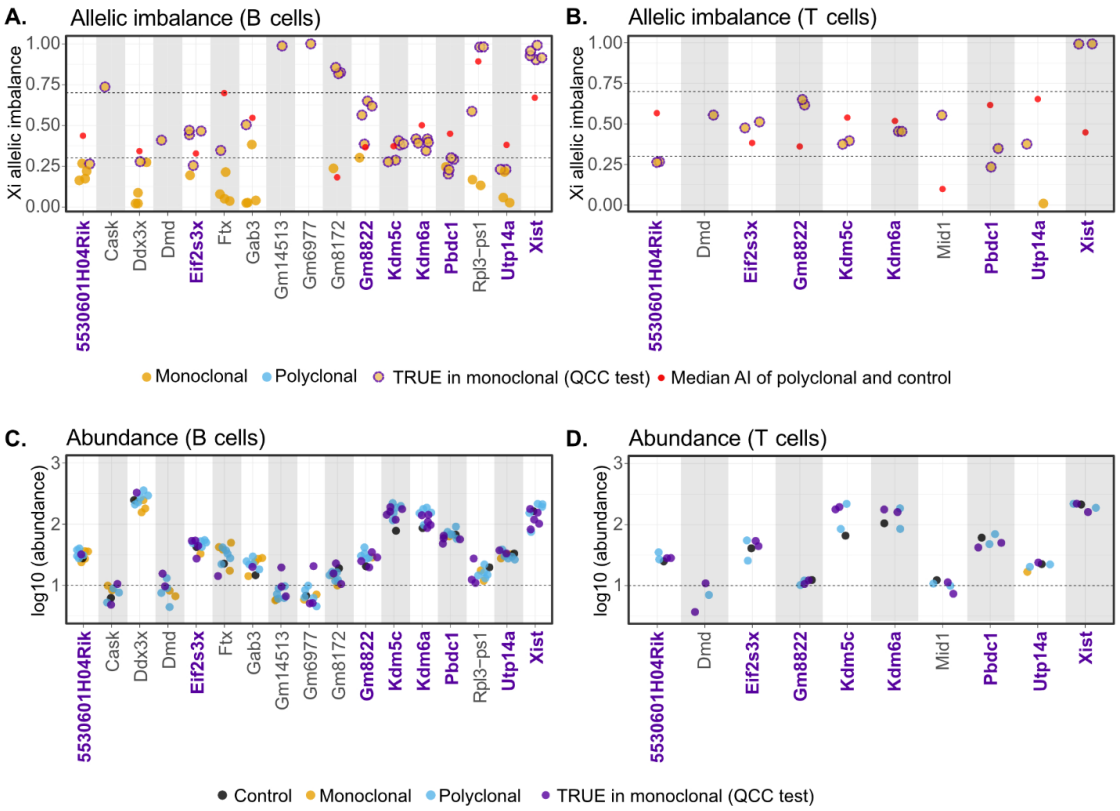


Figure 4. 21. The strategy for identification of X chromosome inactivation escapees. (A) and (B) Allelic imbalance of X-linked genes for B and T cells, respectively. As a convention, *Xi allelic imbalance=1* means that the gene is 100% expressed from the inactive X-linked allele; *Xi allelic imbalance=0* means that only the active X-linked allele was detected. Dots represent genes with expression higher than 10 TMM-normalized counts, and only genes that were statistically different from the sample-corrected threshold at least once are shown. Yellow dots represent monoclonal samples; violet stroke-surrounded yellow dots denote statistical significance for that sample. Red dots represent the median of the allelic imbalance observed for polyclonal and control samples (otherwise excluded from this top panel to compare Xi allelic imbalance of monoclonal samples with the median of polyclonal and control samples). Xi means inactive X chromosome. Statistical significance was calculated by comparing the allelic imbalance with the sample-corrected threshold using binomial test and QCC correction. The threshold was calculated per sample as 0.1 (which is the value usually found in the literature) + the median value of allelic imbalance of all X-linked genes in the sample. (C) and (D) Abundance (TMM-normalized counts) of the same genes and same samples represented in (A), (B). In addition, individual polyclonal and control samples are shown. Violet dots represent the monoclonal samples in which the allelic imbalance significantly deviates from the sample-corrected threshold. Yellow dots represent the other monoclonal samples, blue dots represent the polyclonal samples, and black dots are the control samples. Genes in violet (x-axis) were identified as escapees using the three criteria described in Methods and Results.

Eight XCI escapees were identified in both lymphoid tissues taking into account the used criteria: *5530601H04Rik*, *Eif2s3x*, *Gm8822*, *Kdm5c*, *Kdm6a*, *Pbdc1*, *Utp14a,* and *Xist* (**Figure 4.22 A**). These escapees were plotted along the X chromosome (**Figure 4.22 B**), and it was confirmed that these genes are not clustered (Li et al., 2012). Furthermore, escapees from this study were compared with 117 genes identified as escaping XCI in several studies in different mouse cell lines (Berletch et al., 2015; Li et al., 2012; Wu et al., 2014; Yang et al., 2010) (**Supplementary Table 7.1**). Thirty-six of these escapees were

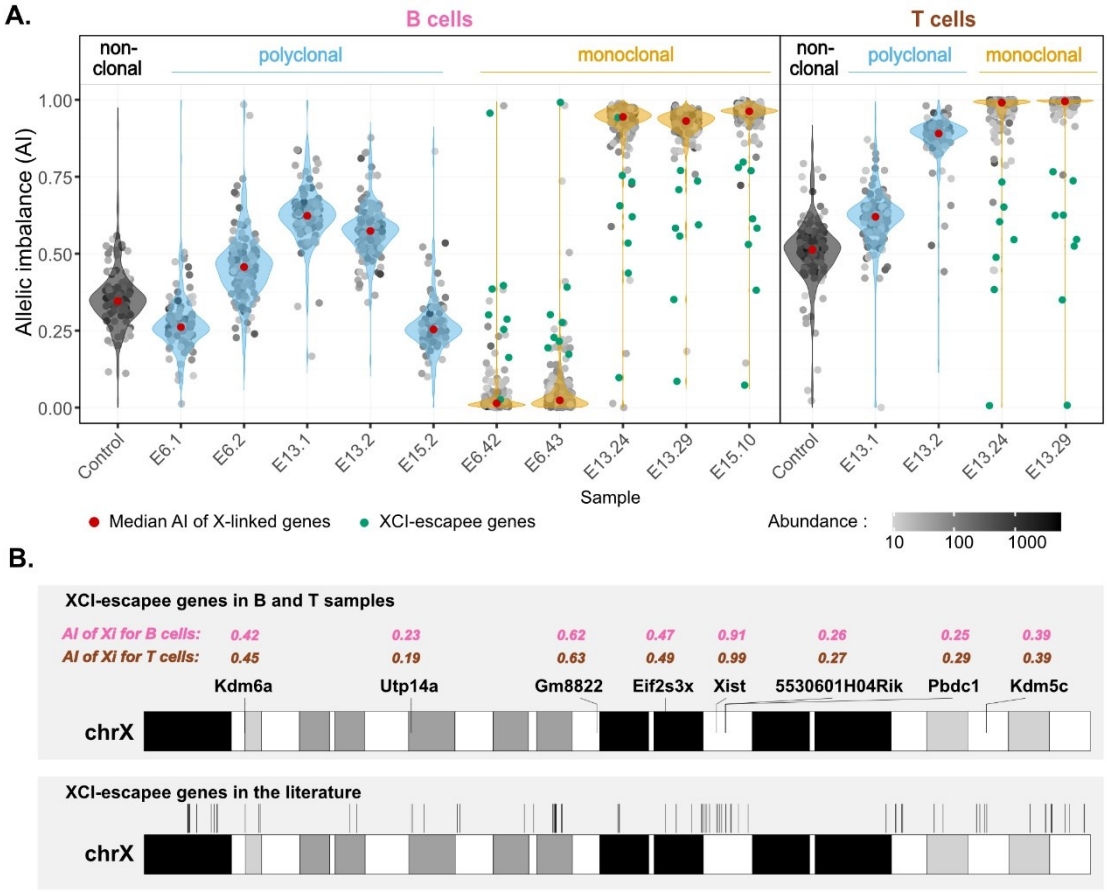


Figure 4. 22. Identification of murine X chromosome inactivation escapees. (A) Distribution of AI values of X-linked genes and identification of XCI escapee genes. Violin plots overlaying dot plots of X-linked genes allelic ratios. For grey dots, the opacity reflects the relative abundance in TMM-normalized counts. Genes significantly escaping XCI (green dots) were identified by comparing the allelic ratio of that gene with a sample-corrected threshold (10% of expression from inactivated X chromosome) and applying the binomial test with QCC correction (Mendelevich et al., 2021). (B) XCI escapee genes on B and T cells annotated along the X chromosome ideogram. The AI values of identified XCI escapee genes are denoted in pink (for B cell samples) and brown (for T cell samples).

excluded from our analysis for lack of expression, two genes by lack of SNPs necessary by default to measure the AI value, and one gene was not found in the annotation reference list used in this work. Of the remaining 78 known escapee genes, 71 genes were not identified as escapees in this study of B and T cells. However, seven of the eight genes identified here belong to this group of known escapees, except for *Gm8822*, a pseudogene that was not detected or reported in the published.

Usually, XCI escapees are identified by one of three systems: 1) single-cell RNA-seq (Borensztein et al., 2017; Chen et al., 2016); 2) heterozygous female mice knockout for specific X-linked genes (Berletch et al., 2015; Yang et al., 2010) or for an X-linked genes fused to a reporter (Wu et al., 2014); 3) and clonal female F1 hybrid cell lines (Calabrese et al., 2012; Li et al., 2012; Splinter et al., 2011). Overall, we conclude that our single-HSC reconstitution approach is a valid new approach to identify XCI escapees *in vivo* in any hematopoietic cell lineage.

4.2.5. Identification of genes with differential AI between B and T cells

As RNA-seq data were produced for the B and T lymphoid lineages, it was interesting to analyze differences between these cells, which have a common lymphoid progenitor but then engage in different development programs. Visualization of AI in a low-dimensional space with the t-SNE algorithm produced two independent clusters for B and T cells based on their differences in allele-specific expression (**Figure 4.23 A**). By analyzing differences in gene expression independently of AI, 5,289 downregulated genes and 4,775 upregulated genes in T cells compared to B cells were found (**Figure 4.23 B**). These differences in tissue-specific gene expression are expected.

A comparison of AI ratios was then carried out between available B and T cells from the same animal (E13.1_B against E13.1_T, E13.2_B against E13.2_T, E13.24_B against E13.24_T, and E13.29_B against E13.29_T) to evaluate if each of the alleles responds uniquely in a different cell type-specific nuclear environment (**Figure 4.24 A**). A high number of genes with significant differential AI was found in each of the four comparisons, defining a group of 146 common genes with a significant differential allele-

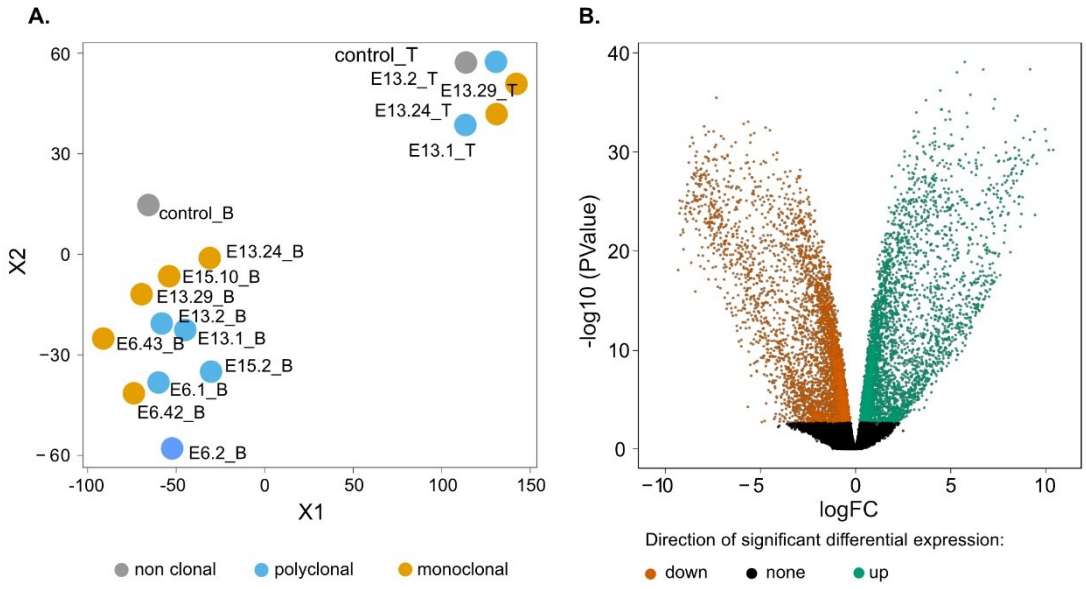


Figure 4. 23. B and T cell type differences in allelic imbalance and gene expression. (A) Visualization of high-dimensional data of autosomal AI for B and cells in a low-dimensional space using (t-SNE algorithm). (B) Volcano plot representing genes with differential expression (upregulated or downregulated) between B and T cells.

specific expression between B and T cells. Since the same genes with different AI values between B and T cells are found in the polyclonal and monoclonal animals, the underlying mechanism is genetic. Statistical enrichment analysis of this group of genes was performed with g: Profiler (Peterson et al., 2020; Raudvere et al., 2019) against all known genes in the mouse genome using a hypergeometric test with correction for multiple testing. As a control, a set of genes with the same dimension but without differential AI between B and T cells was used (**Figure 4.24 B**). Over-representation of genes was found for different functional terms for both set of genes, suggesting that tissue-differential AI is not associated with any particular function.

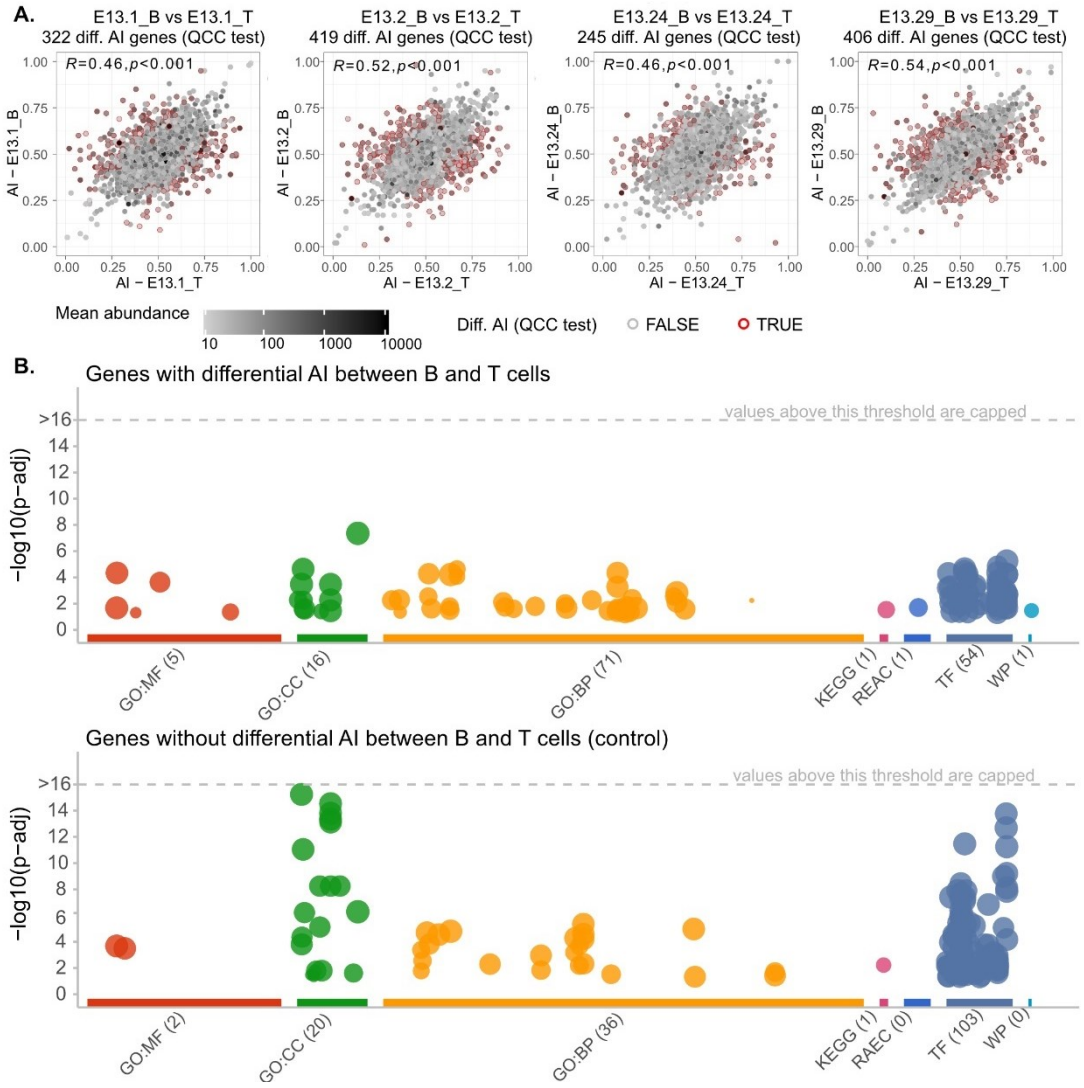


Figure 4. 24. Differences in the B and T cellular environment lead to drastic differences in AI. (A) Dot plots of pairwise comparison of AI between B and T cells within each sample of experiment 13. Red circles signal the genes for which differential AI remained statistically significant after QCC correction on the binomial test. The total number of these genes per comparison is shown above each plot. The Pearson's coefficient correlation for all AI pairwise comparisons is also shown at each dot plot's upper left corner. Mean abundance levels (mean TMM-normalized counts) are continuous greyscale colors. (B) A Manhattan-like plot represents enrichment analysis of genes with significant differential AI found between B and T cells. As a control, a set of genes with the same dimension (146) was randomly sampled from genes without differential AI between B and T cells. Only statistically significant results are shown (hypergeometric test with set counts and size default correction for multiple testing). The x-axis represents the functional terms with the number of overrepresented genes: GO:MF (Gene Ontology: Molecular Function), GO:CC (Gene Ontology: Cellular Component), GO:BP (Gene Ontology: Biological Process), KEGG (Kyoto Encyclopedia of Genes and Genomes), REAC (Reactome), TF (TRANSFAC), and WP (WikiPathways). The y-axis shows the adjusted p-values on the negative log10 scale. Every circle is one term, and the circle size corresponds to the term size (number of genes associated to the term).

5. Discussion and conclusions

Mammalian genomes are diploid, and it is usually assumed that both alleles are equally expressed; differential gene expression between cell types or conditions receives much more attention than the expression differences between the alleles of a given gene. Fifteen years ago, it was demonstrated that RMAE is widespread in mammalian genomes (Gimelbrant et al., 2007) but the prevalence of RMAE *in vivo* and its stability in clones remain controversial (Reinius and Sandberg, 2018; Rv et al., 2021; Vigneau et al., 2018). Furthermore, the parallels between RMAE and XCI seem to be under a vague consensus and have not been challenged until recently (Barreto et al., 2021). RMAE has been well studied *in vitro,* in the human and murine cell lines (Branciamore et al., 2018; Eckersley-Maslin et al., 2014; Gendrel et al., 2014; Gimelbrant et al., 2007; Jeffries et al., 2012, 2016; Pinter et al., 2015; Zwemer et al., 2012). However, the frequency of this phenomenon *in vivo* is an open question because at the tissue level the contribution from each different clone averages out, and the identification or production of clonal cell populations *in vivo* is technically challenging. It has been argued that chromosomal instability in cell lines with prolonged expansion in culture could contribute to the overestimation of RMAE (Reinius and Sandberg, 2015), but for some genes the RMAE patterns detected *in vitro* were validated *in vivo* (Gendrel et al., 2014; Gimelbrant et al., 2007; Marion-Poll et al., 2021). Additionally, although single-cell RNA-seq analysis could be an excellent alternative to study RMAE *in vivo*, the high noise associated to this approach restricts the detection of allele-specific expression to the genes with the highest expression levels and extreme allelic bias (Kim et al., 2015), which is a limitation because the genes under RMAE tend to be expressed at low levels (Eckersley-Maslin et al., 2014; Gendrel et al., 2014; Gimelbrant et al., 2007; Jeffries et al., 2012; Li et al., 2012; Reinius et al., 2016). To answer the question of whether RMAE is present *in vivo* and is clonally stable, we performed a transcriptome-wide analysis of lymphoid cells derived from a single HSC injected into a mouse. In this study, we included technical replicates for RNA-seq libraries of bulk experiments to estimate overdispersion and measured the differential allelic imbalance between clonal populations with high precision (Mendelevich et al., 2021).

Single HSC reconstitutions

The HSC population is highly heterogenous and constitutes a minuscule fraction of bone marrow cells (0.004–0.007%) (Kiel et al., 2007; Uchida et al., 2003; Yang et al., 2005). Even the most refined HSC purification protocols can isolate HSCs at only 40%–50% purity, which means that most of these cells are not authentic LT-HSCs that repopulate an irradiated mouse at a single cell level with long-term and multilineage reconstitution (Wilson et al., 2015). We used three different protocols for HSCs isolation (and one of the protocols with two different approaches), and we ended up selecting protocol 2.1 (LSK $CD48^-$ $CD150^+$ (Kiel et al., 2005)), which provided the highest frequency of reconstituted animals, namely in experiment 6 (19%). The average of all reconstitutions was 7.7%, which is close to the 18-22% range described in different studies (Boyer et al., 2019; Osawa et al., 1996; Wagers et al., 2002). The low reconstitution efficiency could be explained by several reasons. First, the fraction of isolated HSCs with developed protocols is not completely pure. Second, only some of the injected real HSCs successfully migrate through the blood flow and engraft the right niches of the bone marrow. Third, even if the engraftment happens, a single injected cell needs to produce at least 10^9 mature cells to be detectable in the recipient mouse system, such as the hematopoietic organs and blood (Benveniste et al., 2010). Moreover, the long period the cells remain outside the animal, the sorting procedures, and additional handling manipulations could decrease the transplantation efficiency.

The key factor for the success of this work is monoclonality. Female animals and XCI were used to produce an internal control that could confirm the monoclonality. Early in development, a female cell randomly inactivates one of two X chromosomes, and this choice is clonally propagated through cell division and differentiation. By injecting one HSC that has already decided on the choice of XCI, the derived clonal cell population will have an extremely biased AI of X-linked genes because all cells will express the same X chromosome. This is a necessary condition to claim that the hematopoietic system is monoclonal, but it is not a sufficient one because the recipient animal may have received two HSCs that inactivated the same X chromosome. However, given the single cell-sorting protocol, the inspection of the Terasaki plate wells under the microscope, and

the extreme care during the single-cell injections, which were always performed by two individuals and moving the injected mouse to a different cage to avoid injecting the same animal twice, the probability of having animals with two donor HSCs is low. Moreover, the low frequency of animals reconstituted with a single cell reinforces the conclusion that the five animals we studied had a monoclonal hematopoietic system (donor cells) because, even if two cells were introduced, the probability of both surviving, home to the hematopoietic niches, self-renew, and differentiate to produce a long-term multilineage population should be extremely low. This was directly observed in experiment 5, in which two cells were purposedly injected in the same recipient animal (GFP^+ and GFP^- cells), because no reconstituted animal was found with contributions in the blood from two donor HSCs. In a worst-case scenario where the five animals assumed to be monoclonal were in fact reconstituted with two HSCs, the frequency of genes under RMAE would be underestimated only by 50% (similar to XCI, the probability of receiving two cells with opposite skewness), which would not affect our conclusions that RMAE *in vivo* in highly expanded and differentiated hematopoietic cells is a rare event. Genetic labeling of original HSCs with different barcodes is potentially a powerful tool to analyze the monoclonality of samples, but this would involve additional manipulation of the HSCs before transplantation, which could raise several concerns, such as the HSC activation during gene transfers and alteration of its function (Verovskaya et al., 2014).

The deviation from the 1:1 XCI ratio observed in our polyclonal samples and even in unmanipulated animals could be explained by the tissue-specific XCI skewing observed in lymphoblastoid cell lines and whole-blood samples and multiple hematopoietic cell types from two homozygotic twins, while skin and fat tissues showed more balanced skewing (Zito et al., 2019).

It is assumed that choosing which X chromosome to be inactivated during XCI is random; however, this inactivation in F1 hybrid mouse in crosses between classical inbred strains has been described as skewed. The *Xce* (X-controlling element) *locus* genetically described more than 50 years ago by Cattanach interferes in this process (Cattanach, 1970). In mice homozygous for *Xce*, the probability of either parental X chromosome

being inactivated is equal, while in animals heterozygous for *Xce* one parental X chromosome is preferentially inactivated. Different *Xce* alleles have been identified in several inbred mouse strains, including Xce^{a} (CBA/H, C3H/HeH, and BALB/cH), Xce^{b} (C57BL/6H and DBA/2H), and Xce^{c}-like (CAST/Ei), which have different relative strength ($Xce^{a} < Xce^{b} < Xce^{c}$), i.e., the X carrying the stronger allele has the highest probability of being activated (Cattanach and Rasberry, 1994; Cattanach and Williams, 1972; Cattanach et al., 1969; Johnston and Cattanach, 1981; West and Chapman, 1978). Interestingly, we did not observe this AI skewing of X-linked genes to the strongest Xce^{c} allele carried by the CAST mouse. In polyclonal and unmanipulated animals, the AI values are not biased toward zero (CAST, paternal allele). If the animals were injected with two cells, there are three possibilities of reconstitutions: $Xce^{b} + Xce^{b}$ or $Xce^{c} + Xce^{c}$ or $Xce^{b} + Xce^{c}$. This means that we should observe more examples with active X chromosome from CAST, but we observe three of five animals with maternal B6 X chromosome, the one that is more often silenced during XCI, which is one more observation consistent with the conclusion that the probability of having non-monoclonal animals in the group of putative monoclonal animals is low. The two possible explanations for the absence of X chromosome skewness toward CAST X chromosome in our F1 hybrid mice are tissue-specific XCI skewness (Zito et al., 2019), and a noncanonical mechanism for XCI maintenance in lymphocytes (see below) (Sierra and Anguera, 2019).

Regarding the composition of the monoclonal samples, we only used for RNA-seq cells derived from a single HSC that gave rise to long-term reconstitutions, with multilineage (production of myeloid and lymphoid lineages) and self-renewal capacities (reconstitution of secondary recipient animals).

Additionally, an analysis of the V(D)J repertoire was performed to quantitatively address the complexity of the B cell populations in the monoclonal samples. V(D)J rearrangements occur early, in pro-B and pro-T cells, which means that to generate a complex repertoire of clonotypes, a single HSC should significantly expand, producing enough cells before V(D)J recombination is activated during lineage commitment. It was confirmed that monoclonal samples generated a repertoire of V(D)J clonotypes similar to those of the polyclonal and unmanipulated control animals. We conclude that cells

used in single-cell reconstitutions meet the definition of LT-HSC (Dykstra et al., 2007; Kiel et al., 2005; Wilkinson et al., 2020) and that the clonal complexity of B and T cell populations that emerged from single HSC is equivalent to the clonal complexity found in unmanipulated and polyclonal hematopoietic systems.

Identification of autosomal allele-specific expression

Pairwise comparison of differential AI between all samples did not reveal genes under RMAE. This analysis is more suitable to distinguish between samples with many and few genes with AI. However, the study of AI dispersion in B cells between monoclonal and polyclonal sets of animals unmasked 14 genes, which are more allelically imbalanced in the set of monoclonal samples than in the polyclonal samples. This means that the B cells from the different monoclonal animals expanded in a clonal way from an individual HSC and, although they underwent the same differentiation program, each clone revealed unique allele-specific stable transcriptional states. Additionally, by exploring T cells originating from the same HSC, we confirmed that this observation is biologically meaningful and not the result of statistical flukes due to the many comparisons (number of genes) we explored. The AI values of 14 genes with RMAE are highly correlated in B and T cells, meaning that allele-specific expression patterns established prior to CLP state and already present in HSC are independently propagated in the B and T cell lineages.

In our collection of 14 genes, the most solid example of RMAE is *Pkp3* (plakophilin 3, participates in linking cadherins to intermediate filaments in the cytoskeleton), which can be expressed from paternal or maternal or from both alleles randomly. This gene is more skewed for the CAST allele-specific expression in two monoclonal samples and the B6 in two other monoclonal samples, and shows balanced expression from both alleles in one sample. For different reasons, another curious example is the *Igkv6-25* gene. V(D)J rearrangement occurs in pre-B cells. Before this stage, cells expand vigorously and then each undergoes random recombination in one of the two antigen receptor alleles. It has been shown that there are no epigenetic marks in the HSCs pre-determining which

allele undergoes rearrangement first (Alves-Pereira et al., 2014; Farago et al., 2012), and it is unlikely that such mark is present in the CLP (Alves-Pereira et al., 2014). The appearance of the *Igkv6-25* gene in our study could be due to the low frequency of recombination (<1%) found in this gene (Aoki-Ota et al., 2012). If only one or some cells recombined the same allele, the emerging subclone(s) could produce a "RMAE" pattern. Many immunoglobulin genes exist like this one, but the STAR tool used is not the most suitable to detect rearranged immunoglobulins. As these genes are recombined, and sections need to be aligned, this analysis is not tailored for the immunoglobulin genes. However, the fact that only one gene out of the dozens of immunoglobulin genes is identified as having a "RMAE" pattern corroborates prior findings indicating that the immunoglobulin genes are not predetermined at the HSC to rearrange one allele before the other (Alves-Pereira et al., 2014; Farago et al., 2012).

One important question, perhaps the crucial one, is whether the rare stable transcriptional patterns we identified are due to epigenetic marks or somatic genetic variations. The data revealed no obvious LOH events in the lineage differentiation from the HSCs to the CLP for any genes involved. Additionally, bootstrapping analysis of the transcriptomics and exome sequencing data was performed to test how unique are the 14 genes identified with stable allele-specific expression in terms of the difference between AI from RNA and DNA data. If the RMAE data reflects epigenetic biases and not LOH events, the mean difference between DNA (exome sequencing data) and RNA AI values should be higher for these genes than for most random groups of 14 genes randomly sampled from our data, which was the case. Moreover, a whole-genome sequencing study was recently performed using *in vitro* small clones (~500 cells) derived from the HSC of an 8-month-old B6 mouse and cultured for up to 14 days. In this study, about 110 single nucleotide variants and 26 insertions or deletions were uncovered per HSC. The distribution of these mutations was mostly intergenic, with more than 98.5% not falling in exons, and most are not expected to result in transcription changes (Druce, 2021). Back of the envelope calculations based on these frequencies of genetic alterations strongly suggest that the RMAE pattern observed for *Pkp3* could not result from somatic SNPs, insertions or deletions (data not shown) and the authors of the HSC

whole-genome sequencing study have not found mutations in any of the 14 genes we identified as having a RMAE pattern (Michael Milson, personal communication). In addition, these 14 genes are not associated with known common replication fragile sites (Durkin and Glover, 2007; Ma et al., 2012) and do not present molecular features usually associated with these regions, such as high expression levels and large size (Barlow et al., 2013; Helmrich et al., 2006). Taken these results and observations together, somatic genetic variations are unlikely to produce the stable allele-specific expression we report. However, in the case of genes for which the RMAE relies mostly on one monoclonal sample, somatic gene alterations remain a strong possibility. Only future work with epidrugs that interfere with epigenetic marks will allow us to produce conclusive data about these rare putative epigenetic marks.

To uncover if the low number of genes with allele-specific expression patterns identified in HSC clones that underwent extensive proliferation and differentiation steps *in vivo* is due to the conservative methodology we used (Mendelevich et al., 2021) and to produce the most valid comparison possible between *in vivo* and *in vitro* data, the same analysis pipeline was applied to data from Abelson clones. Even after controlling for the methodology, we identified a significantly higher number of genes with allele-specific expression in clones expanded *in vitro* without differentiation. To explain this discrepancy in numbers between cells expanding and differentiating *in vivo* and cells expanding and not differentiating *in vitro*, including Abelson clones and other cells studied in different works (Eckersley-Maslin et al., 2014; Gendrel et al., 2014; Gimelbrant et al., 2007; Zwemer et al., 2012), we propose a model in which the evolutionary selection pressure that promotes phenotypic diversity at the cellular level by regulating RMAE does not always require the absolute clonal stability of allelic biases through the cellular lifespan (**Figure 5.1**). This clonal stability is found in the two most known examples of RMAE necessary to create phenotypic diversity within initially isogenic cell populations: antigen receptor and OR genes. Stable phenotypic diversity is crucial for these cells. In the case of antigen receptor genes, a negative feedback stops V(D)J recombination if the first recombination event involving one of the alleles was successful, preventing the expression of both antigen receptors. After this lock, the cell

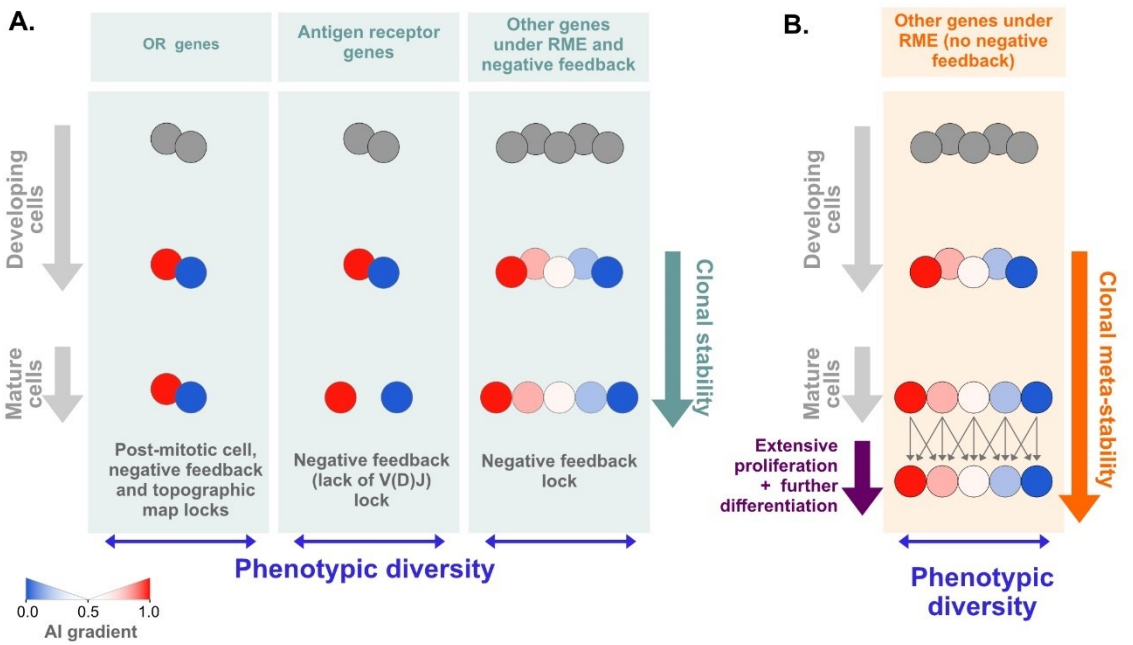


Figure 5. 1. Models of RMAE. (A) For most autosomal genes under RME, the epigenetic states leading to allelic biases are established *de novo* during differentiation and shortly before the genes are expressed. This model of RMAE is characterized by documented (e.g., olfactory receptor and antigen receptor genes) or probable clonal stability due to the existence of locks that stabilize the allelic imbalance (reviewed in (Barreto et al., 2021)). One notable lock is the negative feedback triggered by the protein expression of one allelic form that prevents further gene or allelic activation (or recombination, in the case of the antigen receptors). (B) A model of RMAE in which the allelic imbalance for each clone is meta-stable, i.e., it can change from one cell stage to the other within a certain range during extensive periods of proliferation and differentiation until reaching a new clonal meta-stability. Because of these shifts, the allelic imbalance becomes intraclonally undetectable but is stable within each cell stage and ensures phenotypic diversity. Assuming that HSCs have an initial percentage of genes under RME close to that estimated for cells from collections of developmentally frozen clones grown *in vitro*, our data are compatible with a meta-stable model of RMAE.

cannot rearrange and reinvent its antigen receptor (except in the cases of the receptor editing of the light chain genes and somatic hypermutation during the germinal center reaction). The monoallelic expression of olfactory genes in odorant cells is also stable throughout life and regulated by a negative feedback preventing the expression of more than one allele/gene, which is necessary to preserve the olfactory topographic map. Antigen and olfactory receptors are two compelling but exceptional cases of monoallelic

expression that may not be particularly relevant for other genes under RMAE (**Figure 5. A**). Instead of propagating allelic biases established early throughout many cell divisions, the epigenetic marks could simply emerge before the developmental stage where allele-specific expression is relevant for the cell population phenotype. Cells could change allelic biases in a stochastic way from one stage of differentiation, thus ensuring phenotypic diversity at any developmental stage even if clones are not stable (**Figure 5. B**). We propose that in clones undergoing differentiation, allele-specific patterns are meta-stable, and there is erasure and intraclonal reestablishment of these patterns.

Future work

In the future, to confirm this hypothesis, we plan to compare *in vivo* clones extensively expanded and differentiated from HSC with the same cell-type clones but that did not undergo differentiation. This can be done by sorting cells from the same animals and shortly expand them *ex vivo* without major differentiation to freeze allele-specific marks. Ideally, the experiment would include clones of B cells and clones of HSCs shortly expanded *ex vivo*. We expect to see more allelic biases in these *ex vivo* clones than in our collection of monoclonal B cells that emerged from single HSCs. These *ex vivo* clones should reveal allelic biases present in a specific developmental stage, which would be masked in our in *vivo* experiment because of ongoing differentiation. It would also be interesting to study the change of allelic biases and lineage development and connect these changes to phenotypic diversity.

X chromosome inactivation

One of the most extreme examples of monoallelic expression is XCI. This feature is traditionally used to identify clonally derived cells. At the cellular level, the normal 50:50 ratio of the cells expressing the maternal or paternal X chromosome found in polyclonal populations is drastically skewed in clonal and oligoclonal samples. However, XCI is limited to one sex, it is skewed in female hematopoietic cells, and it quickly becomes uninformative as the number of stem cells increases. In addition, this skewing increases

with age, leading to low resolution in clonal assays (Ayachi et al., 2020). Identifying autosomal regions by focusing on polymorphisms with stable epigenetic allele-specific patterns in hematopoietic lineages could be used to develop assays that estimate the clonal structure of the hematopoietic system. Moreover, these assays based on autosomal epigenetic patterns could also be applied to male cells.

One of the X chromosomes in female cells is inactive due to XCI. However, genes can escape from this mechanism. In mice, XCI escapees have been studied using three systems: 1) single-cell RNA-seq (Borensztein et al., 2017; Chen et al., 2016); 2) heterozygous female mice knockout for specific X-linked genes, such as *Xist* or *Hprt* (Berletch et al., 2015; Yang et al., 2010) or heterozygous female mice for an X-linked gene linked to a reporter (Wu et al., 2014); 3) and clonal female F1 hybrid cell lines (Calabrese et al., 2012; Li et al., 2012; Splinter et al., 2011). In the first method, technical noise leads to a high number of false positives (Kim et al., 2015). The second method is performed on an animal model and involves a genetically engineered *Xist locus*; in the case of knockout mice, the activation of one allele is imposed by the deletion of *Xist*. The third method is based on *in vitro* systems. In short, all three systems have shortcomings. Here we propose an approach to study lineage-specific XCI *in vivo* using genetically unmanipulated cells. We identified seven (*5530601H04Rik, Eif2s3x, Kdm5c, Kdm6a, Pbdc1, Utp14a*, and *Xist*) XCI escapees in B and T cells previously found in different tissues (Berletch et al., 2015; Li et al., 2012; Wu et al., 2014; Yang et al., 2010). A set of known escapees were not identified in this study, which could be due to the tissue-specificity of X escapees (Berletch et al., 2015) or the restrictive criteria used in this work. Overall, we show that single HSC reconstitution is an effective method to study lineage-specific XCI in blood.

Traditionally, the mechanism for XCI maintenance is thought to be the same across all somatic cells. However, this model was drawn based on studies performed with immortalized or primary fibroblasts, cancer cell lines, differentiated stem cells, and neural precursor cells (Sierra and Anguera, 2019). In 2006, a study on murine lymphocytes reported that XCI maintenance could be noncanonical in hematopoietic cells. The inactive X chromosome was characterized by *Xist* RNA FISH and H3K27me3

staining from sorted female mouse hematopoietic cells in this study. Unexpectedly, it was found that *Xist* RNA clusters and H3K27me3 *foci* are present in LSKs and lymphoid progenitors, but the H3K27me3 signal was lost in committed cells downstream of the lymphoid progenitors. The *Xist* RNA cluster was detected in pro-B and pro-T cells, and then it was significantly decreased in pre-B and pre-T cells and detected again in a significant fraction of mature B and T cells. Allele-specific RT-PCR analysis of *Xist* and four X-linked genes in pre-B and pre-T, and mature B and T cells indicated that these genes were not reactivated from the inactive X chromosome, suggesting that dosage compensation was still preserved (Savarese et al., 2006). Subsequently, a series of publications by Anguera et al. revealed that the mechanism of hematopoietic XCI maintenance in lymphocytes is dynamic and differs from that of other somatic cells. By analyzing mature naïve B and T cells isolated from mice and humans, it was found that these lymphocytes lack typical heterochromatin marks and the *Xist* RNA cloud. *In vitro* activation of both lymphocytes triggered the return of *Xist* RNA signal and some chromatin marks to the inactive X chromosome, but with kinetic differences between B and T cells. *Xist* was continuously expressed from naïve and activated lymphocytes (Wang et al., 2016). Next, the dynamics of XCI maintenance of B and T cells during differentiation was studied in more detail. In HSCs and CLPs, *Xist* RNA clouds and the H3K27me3 mark were canonical and similar to those of fibroblasts. Then, in B cell development, *Xist* RNA disappears from the inactive X chromosome in pro-B, pre-B, and immature cells, whereas heterochromatin marks are still present in pro-B cells and progressively decrease during development (Syrett et al., 2017). In the T cell lineage, *Xist* RNA disappears in the DN1 stage, re-localizes in the DN2 and DN3 stages, and then disappears again from the inactive X chromosome in the DN4 stage, with a high correlation to the presence of the H3K27me3 mark (Syrett et al., 2019). The functional importance for dynamic localization of *Xist* RNA at inactive X chromosome it still unknown. Moreover, the *CXCR3*, *CD40L,* and *TLR7* genes, which are immune-related X-linked genes, were found to be biallelically expressed in lymphocytes of females (XX) and males with Klinefelter Syndrome (XXY, with one inactive X chromosome resulting from XCI), both prone to develop SLE, an autoimmune disease. B cell lines of patients with SLE also displayed more cells with biallelic expression of these genes than control

cell lines (Souyris et al., 2018; Wang et al., 2016). The escape from XCI of immune-related X-linked genes and its connection to autoimmune disease and how other X-linked genes escape from dynamic XCI the during different lymphocyte differentiation stages are topics worthy of further investigation. The clonally derived hematopoietic system *in vivo* used in this work could be applied to tackle these questions, as well as tissue-specific XCI skewing. Additionally, immunodeficient mice could be single-cell reconstituted with human HSCs to evaluate XCI in the human hematopoietic system (Beyer and Muench, 2017).

Autosomal versus XCI parallels

The precise mechanism or mechanisms underlying RMAE is until now unknown. Due to the common features between XCI and RMAE, namely the randomness and clonal propagation, parallels between these two classes of monoallelic expression are typically drawn when discussing RMAE (Chess, 2016; Gendrel et al., 2016; Goldmit and Bergman, 2004; Mostoslavsky et al., 2001; Pereira et al., 2003). At least one gene, *Smchd1*, was suggested to regulate XCI of X-linked genes (completion or maintenance) and genes with RMAE. It was found that dramatically reduced expression of the this gene alters the expression of the clustered protocadherins in the brain (Mould et al., 2013) and killer cell lectin-like receptors subfamily A (also known as *Ly49*) in transformed mouse embryonic fibroblasts and tumors (Leong et al., 2013); both genes are monoallelically expressed (Esumi et al., 2005; Held et al., 1995). LINE-1 participate in XCI by spreading the silencing regions (Chow et al., 2010). Similarly, these elements were also found to flank genes under RMAE with high density, leading to the proposal that LINE-1 could also participate in the monoallelic expression of these genes (Allen et al., 2003). Finally, two human non-coding RNA autosomal genes, *ASAR6* and *ASAR15*, share some characteristics with *Xist*. Briefly, they are randomly and monoallelically expressed from the later replicating allele, replicate asynchronously in coordination with other linked monoallelic genes, remain associated with the chromosome from which they are expressed, contain a high density of LINE-1, and their deletion results in late replication and activation of the previously silent alleles of nearby genes (Donley et al., 2013, 2015;

Stoffregen et al., 2011). Despite these similarities, here we observed that, after extensive proliferation and differentiation *in vivo,* the regions in the autosomal chromosomes behaving like the X chromosomes in terms of the stable transcriptional states may not exist or represent only an extremely small fraction of the genome. We suggest that RMAE lacks the stability seen in XCI, which depends on a multilayered process of silencing (Dossin and Heard, 2021). It is probable that a XCI-like stability applied to the autosomes would compromise the dynamics necessary for the different programs of hematopoietic development.

Identification of genes with differential AI between B and T cells

Monoallelic expression can result from epigenetic or genetic mechanisms. Allele-specific expression due to genetic features depends on the underlying DNA sequence around gene regulatory regions. On the other hand, in gene expression resulting from epigenetic features, the alleles can have identical DNA sequences, be present in the same cell, and undergo the same *trans*-acting environment, and still present differences in chromatin composition, bound transcription factors, and expression status (Gibney and Nolan, 2010). The differences found in allele-specific expression between B and T cells are surprising. However, given that the differences are present in the monoclonal and polyclonal samples, they have mainly a genetic component, with the B6 and CAST alleles responding differently to the distinct nuclear environment of B and T cells. Therefore, it would be interesting to investigate how the early evolution of gene-regulatory sequences (C57BL/6J and Cast/EiJ diverged about 1 million years ago (Wade et al., 2002)) impacts the allelic biases observed in the different cell types from the same animal. We did not reveal any relevant function associated to genes with differential AI between B and T cells; for instance, the high enrichment in transcription factor binding sites was equally found for a control set of genes without allelic biases. The genes with alleles that respond differently to the B and T cell environment probably carry different mutations in the regulatory regions of these alleles, but the work of mapping these mutations needs to be done.

Conclusion

This is the first report of an allele-specific genome-wide transcriptome analysis of lymphoid cell populations derived from a single HSC *in vivo* after extensive differentiation. It is here assumed that RMAE patterns are established during the differentiation of embryonic stem cells (Eckersley-Maslin et al., 2014; Gendrel et al., 2014; Marion-Poll et al., 2021) and we tested whether these patterns, once established in HSCs, are stably maintained across subsequent differentiation steps (Gendrel et al., 2014; Marion-Poll et al., 2021).

Heterogeneous repopulation phenotypes revealed by single HSC transplantation assays suggest that they are caused by epigenetic mechanisms present in HSCs (Benveniste et al., 2010; Dykstra et al., 2007; Morita et al., 2010; Sieburg et al., 2006; Yu et al., 2016). Despite being "stem cells", HSCs are certainly more differentiated than embryonic stem cells, which are known to display RMAE patterns (Eckersley-Maslin et al., 2014) and this cell is less developed than the HSC. Thus, it is reasonable to postulate that HSCs also have genes under RMAE. Indeed, in our designed approach, where we studied clonally differentiated lymphoid cells from a single HSC, we found genes that maintain these allelic biases, meaning that these marks should already be present in the original HSC. However, we found that the percentage of genes with RMAE in the monoclonal hematopoietic system in which cells suffered prolonged and extensive cell expansion and differentiation is much lower (<0.2%) than the estimates obtained from frozen cell clones *in vitro* without undergoing extensive development, suggesting that these stable allele-specific transcriptional patterns are metastable and could be erased and reestablished during lineage commitment. This means that RMAE is not stably propagated during differentiation, unlike XCI.

Additionally, we tested the approach used here to identify XCI escapees. We showed that the single HSC reconstitution is an alternative approach with several advantages over the systems that have been used. Our approach allows the study of XCI *in vivo* without genetic manipulations and using bulk RNA-seq instead of the noisy single-cell RNA-seq.

6. Bibliographic references

Adegbola, A.A., Cox, G.F., Bradshaw, E.M., Hafler, D.A., Gimelbrant, A., and Chess, A. (2015). Monoallelic expression of the human FOXP2 speech gene. Proc. Natl. Acad. Sci. *112*, 6848–6854.

Adolfsson, J., Borge, O.J., Bryder, D., Theilgaard-Mönch, K., Åstrand-Grundström, I., Sitnicka, E., Sasaki, Y., and Jacobsen, S.E.W. (2001). Upregulation of Flt3 expression within the bone marrow Lin-Sca1+c-kit+ stem cell compartment is accompanied by loss of self-renewal capacity. Immunity *15*, 659–669.

Aifantis, I., Buer, J., Von Boehmer, H., and Azogui, O. (1997). Essential role of the pre-T cell receptor in allelic exclusion of the T cell receptor β locus. Immunity *7*, 601–607.

Akashi, K., Traver, D., Miyamoto, T., and Weissman, I.L. (2000). A clonogenic common myeloid progenitor that gives rise to all myeloid lineages. Nature *404*, 193–197.

Allen, E., Horvath, S., Tong, F., Kraft, P., Spiteri, E., Riggs, A.D., and Marahrens, Y. (2003). High concentrations of long interspersed nuclear element sequence distinguish monoallelically expressed genes. Proc. Natl. Acad. Sci. U. S. A. *100*, 9940–9945.

Alt, F.W., Yancopoulos, G.D., Blackwell, T.K., Wood, C., Thomas, E., Boss, M., Coffman, R., Rosenberg, N., Tonegawa, S., and Baltimore, D. (1984). Ordered rearrangement of immunoglobulin heavy chain variable region segments. EMBO J. *3*, 1209–1219.

Alves-Pereira, C.F., De Freitas, R., Lopes, T., Gardner, R., Marta, F., Vieira, P., and Barreto, V.M. (2014). Independent recruitment of Igh alleles in V(D)J recombination. Nat. Commun. *5*, 5623–15.

Anders, S. (2010). Babraham Bioinformatics - FastQC A Quality Control tool for High Throughput Sequence Data.

Anders, S., and Huber, W. (2010). Differential expression analysis for sequence count data. Genome Biol. *11*, 1–12.

Anderson, S.K. (2014). Probabilistic bidirectional promoter switches: Noncoding RNA takes control. Mol. Ther. - Nucleic Acids *3*, e191.

Aoki-Ota, M., Torkamani, A., Ota, T., Schork, N., and Nemazee, D. (2012). Skewed

Primary Igκ Repertoire and V–J Joining in C57BL/6 Mice: Implications for Recombination Accessibility and Receptor Editing. J. Immunol. *188*, 2305–2315.

Armelin-Corre, L.M., Gutiyam, L.M., Brandt, D.Y.C., and Malnic, B. (2014). Nuclear compartmentalization of odorant receptor genes. Proc. Natl. Acad. Sci. U. S. A. *111*, 2782–2787.

Arnhold, I.J.P., França, M.M., Carvalho, L.R., Mendonca, B.B., and Jorge, A.A.L. (2015). Role of GLI2 in hypopituitarism phenotype. J. Mol. Endocrinol. *54*, R141–R150.

Aseem, O., Barth, J.L., Klatt, S.C., Smith, B.T., and Argraves, W.S. (2013). Cubilin expression is monoallelic and epigenetically augmented via PPARs. BMC Genomics *14*, 405.

Ayachi, S., Buscarlet, M., and Busque, L. (2020). 60 Years of clonal hematopoiesis research: From X-chromosome inactivation studies to the identification of driver mutations. Exp. Hematol. *83*, 2–11.

Azuara, V., Perry, P., Sauer, S., Spivakov, M., Jørgensen, H.F., John, R.M., Gouti, M., Casanova, M., Warnes, G., Merkenschlager, M., et al. (2006). Chromatin signatures of pluripotent cell lines. Nat. Cell Biol. *8*, 532–538.

Babak, T., Deveale, B., Tsang, E.K., Zhou, Y., Li, X., Smith, K.S., Kukurba, K.R., Zhang, R., Li, J.B., Van Der Kooy, D., et al. (2015). Genetic conflict reflected in tissue-specific maps of genomic imprinting in human and mouse. Nat. Genet. *47*, 544–549.

Baran, Y., Subramaniam, M., Biton, A., Tukiainen, T., Tsang, E.K., Rivas, M.A., Pirinen, M., Gutierrez-Arcelus, M., Smith, K.S., Kukurba, K.R., et al. (2015). The landscape of genomic imprinting across diverse adult human tissues. Genome Res. *25*, 927–936.

Barlow, D.P. (1995). Gametic imprinting in mammals. Science *270*, 1610–1613.

Barlow, J.H., Faryabi, R.B., Callén, E., Wong, N., Malhowski, A., Chen, H.T., Gutierrez-Cruz, G., Sun, H.W., McKinnon, P., Wright, G., et al. (2013). Identification of early replicating fragile sites that contribute to genome instability. Cell *152*, 620–632.

Barreto, V.M., Kubasova, N., Alves-Pereira, C.F., and Gendrel, A.-V. (2021). X-

Chromosome Inactivation and Autosomal Random Monoallelic Expression as “Faux Amis.” Front. Cell Dev. Biol. *9*, 2599.

Barton, S.C., Surani, M.A.H., and Norris, M.L. (1984). Role of paternal and maternal genomes in mouse development. Nature *311*, 374–376.

Basson, C.T., Cowley, G.S., Solomon, S.D., Weissman, B., Poznanski, A.K., Traill, T.A., Seidman, J.G., and Seidman, C.E. (1994). The Clinical and Genetic Spectrum of the Holt-Oram Syndrome (Heart-Hand Syndrome). N. Engl. J. Med. *330*, 885–891.

Basson, C.T., Bachinsky, D.R., Lin, R.C., Levi, T., Elkins, J.A., Soults, J., Grayzel, D., Kroumpouzou, E., Traill, T.A., Leblanc-Straceski, J., et al. (1997). Mutations in human cause limb and cardiac malformation in Holt-Oram syndrome. Nat. Genet. *15*, 30–35.

Becker, K.G., Barnes, K.C., Bright, T.J., and Wang, S.A. (2004). The Genetic Association Database. Nat. Genet. *36*, 431–432.

Benveniste, P., Cantin, C., Hyam, D., and Iscove, N.N. (2003). Hematopoietic stem cells engraft in mice with absolute efficiency. Nat. Immunol. *4*, 708–713.

Benveniste, P., Frelin, C., Janmohamed, S., Barbara, M., Herrington, R., Hyam, D., and Iscove, N.N. (2010). Intermediate-Term Hematopoietic Stem Cells with Extended but Time-Limited Reconstitution Potential. Cell Stem Cell *6*, 48–58.

Berletch, J.B., Ma, W., Yang, F., Shendure, J., Noble, W.S., Disteche, C.M., and Deng, X. (2015). Escape from X Inactivation Varies in Mouse Tissues. PLoS Genet. *11*, e1005079.

Bernstein, B.E., Mikkelsen, T.S., Xie, X., Kamal, M., Huebert, D.J., Cuff, J., Fry, B., Meissner, A., Wernig, M., Plath, K., et al. (2006). A Bivalent Chromatin Structure Marks Key Developmental Genes in Embryonic Stem Cells. Cell *125*, 315–326.

Beyer, A.I., and Muench, M.O. (2017). Comparison of human hematopoietic reconstitution in different strains of immunodeficient mice. Stem Cells Dev. *26*, 102–112.

Bix, M., and Locksley, R.M. (1998). Independent and epigenetics regulation of the interleukin-4 alleles in CD4+ T cells. Science *281*, 1351–1354.

Bolotin, D.A., Poslavsky, S., Mitrophanov, I., Shugay, M., Mamedov, I.Z., Putintseva, E. V., and Chudakov, D.M. (2015). MiXCR: Software for comprehensive adaptive immunity profiling. Nat. Methods *12*, 380–381.

Bolotin, D.A., Poslavsky, S., Davydov, A.N., Frenkel, F.E., Fanchi, L., Zolotareva, O.I., Hemmers, S., Putintseva, E. V, Obraztsova, A.S., Shugay, M., et al. (2017). Antigen receptor repertoire profiling from RNA-seq data HHS Public Access Author manuscript. Nat Biotechnol *35*, 908–911.

Bönnemann, C.G., Cox, G.F., Shapiro, F., Wu, J.J., Feener, C.A., Thompson, T.G., Anthony, D.C., Eyre, D.R., Darras, B.T., and Kunkel, L.M. (2000). A mutation in the alpha 3 chain of type IX collagen causes autosomal dominant multiple epiphyseal dysplasia with mild myopathy. Proc. Natl. Acad. Sci. U. S. A. *97*, 1212–1217.

Borel, C., Ferreira, P.G., Santoni, F., Delaneau, O., Fort, A., Popadin, K.Y., Garieri, M., Falconnet, E., Ribaux, P., Guipponi, M., et al. (2015). Biased allelic expression in human primary fibroblast single cells. Am. J. Hum. Genet. *96*, 70–80.

Borensztein, M., Syx, L., Ancelin, K., Diabangouaya, P., Picard, C., Liu, T., Liang, J. Bin, Vassilev, I., Galupa, R., Servant, N., et al. (2017). Xist-dependent imprinted X inactivation and the early developmental consequences of its failure. Nat. Struct. Mol. Biol. *24*, 226–233.

Boyer, S.W., Rajendiran, S., Beaudin, A.E., Smith-Berdan, S., Muthuswamy, P.K., Perez-Cunningham, J., Martin, E.W., Cheung, C., Tsang, H., Landon, M., et al. (2019). Clonal and Quantitative In Vivo Assessment of Hematopoietic Stem Cell Differentiation Reveals Strong Erythroid Potential of Multipotent Cells. Stem Cell Reports *12*, 801–815.

Branciamore, S., Valo, Z., Li, M., Wang, J., Riggs, A.D., and Singer-Sam, J. (2018). Frequent monoallelic or skewed expression for developmental genes in CNS-derived cells and evidence for balancing selection. Proc. Natl. Acad. Sci. U. S. A. *115*, E10379–E10386.

Buil, A., Brown, A.A., Lappalainen, T., Viñuela, A., Davies, M.N., Zheng, H.., Richards, J.B., Glass, D., Small, K.S., Durbin, R., et al. (2015). Gene-gene and gene-environment interactions detected by transcriptome sequence analysis in twins. Nat. Genet. *47*, 88.

Bullard, J.H., Purdom, E., Hansen, K.D., and Dudoit, S. (2010). Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments. BMC Bioinformatics *11*, 1–13.

Calabrese, J.M., Sun, W., Song, L., Mugford, J.W., Williams, L., Yee, D., Starmer, J., Mieczkowski, P., Crawford, G.E., and Magnuson, T. (2012). Site-specific silencing of regulatory elements as a mechanism of X inactivation. Cell *151*, 951–963.

Calado, D.P., Paixão, T., Holmberg, D., and Haury, M. (2006). Stochastic monoallelic expression of IL-10 in T cells. J. Immunol. *177*, 5358–5364.

Capparelli, R., Costabile, A., Viscardi, M., and Iannelli, D. (2004). Monoallelic expression of mouse Cd4 gene. Mamm. Genome *15*, 579–584.

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

Castel, S.E., Levy-Moonshine, A., Mohammadi, P., Banks, E., and Lappalainen, T. (2015). Tools and best practices for data processing in allelic expression analysis. Genome Biol. *16*, 1–12.

Cattanach, B.M. (1970). Controlling elements in the mouse X-chromosome: III. Influence upon both parts of an X divided by rearrangement. Genet. Res. *16*, 293–301.

Cattanach, B.M., and Rasberry, C. (1994). Identification of the Mus castaneus Xce allele. Mouse Genome *92*.

Cattanach, B.M., and Williams, C.E. (1972). Evidence of non-random X chromosome activity in the mouse. Genet. Res. *19*, 229–240.

Cattanach, B.M., Pollard, C.E., and Perez, J.N. (1969). Controlling elements in the mouse X-chromosome: I. Interaction with the X-linked genes. Genet. Res. *14*, 223–235.

Cebra, J.J., and Goldstein, G. (1965). Chromatographic purification of tetramethylrhodamine-immune globulin conjugates and their use in the cellular localization of rabbit gamma-globulin polypeptide chains. J. Immunol. *95*, 230–245.

Challen, G.A., Boles, N.C., Chambers, S.M., and Goodell, M.A. (2010). Distinct

Hematopoietic Stem Cell Subtypes Are Differentially Regulated by TGF-β1. Cell Stem Cell *6*, 265–278.

Chan, H.W., Kurago, Z.B., Stewart, C.A., Wilson, M.J., Martin, M.P., Mace, B.E., Carrington, M., Trowsdale, J., and Lutz, C.T. (2003). DNA methylation maintains allele-specific KIR gene expression in human natural killer cells. J. Exp. Med. *197*, 245–255.

Chen, G., Schell, J.P., Benitez, J.A., Petropoulos, S., Yilmaz, M., Reinius, B., Alekseenko, Z., Shi, L., Hedlund, E., Lanner, F., et al. (2016). Single-cell analyses of X Chromosome inactivation dynamics and pluripotency during differentiation. Genome Res. *26*, 1342–1354.

Chen, G., Ning, B., and Shi, T. (2019). Single-cell RNA-seq technologies and related computational data analysis. Front. Genet. *10*, 317.

Cheng, H., Zheng, Z., and Cheng, T. (2020). New paradigms on hematopoietic stem cell differentiation. Protein Cell *11*, 34–44.

Chess, A. (2013). Random and non-random monoallelic expression. Neuropsychopharmacology *38*, 55–61.

Chess, A. (2016). Monoallelic Gene Expression in Mammals. Annu. Rev. Genet. *50*, 317–327.

Chess, A., Simon, I., Cedar, H., Axel, R., Barlow, D.P., Stöger, R., Herrmann, B.G., Saito, K., Schweifer, N., Bartolomei, M.S., et al. (1994). Allelic inactivation regulates olfactory receptor gene expression. Cell *78*, 823–834.

Chi, X., Li, Y., and Qiu, X. (2020). V(D)J recombination, somatic hypermutation and class switch recombination of immunoglobulins: mechanism and regulation. Immunology *160*, 233–247.

Chow, J.C., and Heard, E. (2010). Nuclear organization and dosage compensation. Cold Spring Harb. Perspect. Biol. *2*, a000604.

Chow, J.C., Ciaudo, C., Fazzari, M.J., Mise, N., Servant, N., Glass, J.L., Attreed, M., Avner, P., Wutz, A., Barillot, E., et al. (2010). LINE-1 activity in facultative heterochromatin

formation during X chromosome inactivation. Cell *141*, 956–969.

Clowney, E.J., Legros, M.A., Mosley, C.P., Clowney, F.G., Markenskoff-Papadimitriou, E.C., Myllys, M., Barnea, G., Larabell, C.A., and Lomvardas, S. (2012). Nuclear aggregation of olfactory receptor genes governs their monogenic expression. Cell *151*, 724–737.

Cobaleda, C., Jochum, W., and Busslinger, M. (2007). Conversion of mature B cells into T cells by dedifferentiation to uncommitted progenitors. Nature *449*, 473–477.

Conesa, A., Madrigal, P., Tarazona, S., Gomez-Cabrero, D., Cervera, A., McPherson, A., Szcześniak, M.W., Gaffney, D.J., Elo, L.L., Zhang, X., et al. (2016). A survey of best practices for RNA-seq data analysis. Genome Biol. 2016 171 *17*, 1–19.

Costa-Silva, J., Domingues, D., and Lopes, F.M. (2017). RNA-Seq differential expression analysis: An extended review and a software tool. PLoS One *12*, e0190152.

Davies, G.E., Locke, S.M., Wright, P.W., Li, H., Hanson, R.J., Miller, J.S., and Anderson, S.K. (2007). Identification of bidirectional promoters in the human KIR genes. Genes Immun. *8*, 245–253.

Degl'Innocenti, A., and D'Errico, A. (2017). Regulatory features for odorant receptor genes in the mouse genome. Front. Genet. *8*, 19.

Degner, J.F., Marioni, J.C., Pai, A.A., Pickrell, J.K., Nkadori, E., Gilad, Y., and Pritchard, J.K. (2009). Effect of read-mapping biases on detecting allele-specific expression from RNA-sequencing data. Bioinformatics *25*, 3207–3212.

Deng, Q., Ramsköld, D., Reinius, B., and Sandberg, R. (2014). Single-cell RNA-seq reveals dynamic, random monoallelic gene expression in mammalian cells. Science *343*, 193–196.

Dick, J.E. (2003). Self-renewal writ in blood. Nature *423*, 231–233.

Dobin, A., Davis, C.A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., and Gingeras, T.R. (2013). STAR: Ultrafast universal RNA-seq aligner. Bioinformatics *29*, 15–21.

Donley, N., Stoffregen, E.P., Smith, L., Montagna, C., and Thayer, M.J. (2013).

Asynchronous Replication, Mono-Allelic Expression, and Long Range Cis-Effects of ASAR6. PLoS Genet. *9*, e1003423.

Donley, N., Smith, L., and Thayer, M.J. (2015). ASAR15, A cis-Acting Locus that Controls Chromosome-Wide Replication Timing and Stability of Human Chromosome 15. PLoS Genet. *11*, e1004923.

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

Druce, M.C. (2021). The Impact of Ageing, Replication and Stress on Genome Stability in Hematopoietic Stem Cells. PhD Diss. Ruperto Carola Univ. Heidelberg, Ger. 1–157.

Durkin, S.G., and Glover, T.W. (2007). Chromosome fragile sites. Annu. Rev. Genet. *41*, 169–192.

Dykstra, B., Kent, D., Bowie, M., McCaffrey, L., Hamilton, M., Lyons, K., Lee, S.J., Brinkman, R., and Eaves, C. (2007). Long-Term Propagation of Distinct Hematopoietic Differentiation Programs In Vivo. Cell Stem Cell *1*, 218–229.

Eckersley-Maslin, M.A., and Spector, D.L. (2014). Random monoallelic expression: Regulating gene expression one allele at a time. Trends Genet. *30*, 237–244.

Eckersley-Maslin, M.A., Thybert, D., Bergmann, J.H., Marioni, J.C., Flicek, P., and Spector, D.L. (2014). Random monoallelic gene expression increases upon embryonic stem cell differentiation. Dev. Cell *28*, 351–365.

Endo, Y., Sugimura, H., and Kino, I. (1995). Monoclonality of normal human colonic crypts. Pathol. Int. *45*, 602–604.

Engreitz, J.M., Pandya-Jones, A., McDonel, P., Shishkin, A., Sirokman, K., Surka, C., Kadri, S., Xing, J., Goren, A., Lander, E.S., et al. (2013). The Xist lncRNA exploits three-dimensional genome architecture to spread across the X chromosome. Science *341*, 1237973.

Ensminger, A.W., and Chess, A. (2004). Coordinated replication timing of monoallelically expressed genes along human autosomes. Hum. Mol. Genet. *13*, 651–658.

Esumi, S., Kakazu, N., Taguchi, Y., Hirayama, T., Sasaki, A., Hirabayashi, T., Koide, T., Kitsukawa, T., Hamada, S., and Yagi, T. (2005). Monoallelic yet combinatorial expression of variable exons of the protocadherin-α gene cluster in single neurons. Nat. Genet. *37*, 171–176.

Ewels, P., Magnusson, M., Lundin, S., and Käller, M. (2016). MultiQC: Summarize analysis results for multiple tools and samples in a single report. Bioinformatics *32*, 3047–3048.

Farago, M., Rosenbluh, C., Tevlin, M., Fraenkel, S., Schlesinger, S., Masika, H., Gouzman, M., Teng, G., Schatz, D., Rais, Y., et al. (2012). Clonal allelic predetermination of immunoglobulin-κ rearrangement. Nature *490*, 561–565.

Farooqi, M.S., Mishra, D., Chaturvedi, K.K., Rai, A., Lal, S.B., Kumar, S., Bhati, J., and Sharma, A. (2019). A Review on Recent Statistical Models for RNA-Seq Data. J. Appl. Bioinforma. Comput. Biol. *8*, 1–5.

Farooqi, M.S., Kumar, D., Mishra, D.C., Rai, A., and Singh, N.K. (2021). A hybrid method for differentially expressed genes identification and ranking from RNA-Seq data. Int. J. Bioinform. Res. Appl. *17*, 38–52.

Ferguson-Smith, A.C., and Bourc'his, D. (2018). The discovery and importance of genomic imprinting. Elife *7*, e42368.

Frazer, K.A., Eskin, E., Kang, H.M., Bogue, M.A., Hinds, D.A., Beilharz, E.J., Gupta, R. V., Montgomery, J., Morenzoni, M.M., Nilsen, G.B., et al. (2007). A sequence-based variation map of 8.27 million SNPs in inbred mouse strains. Nature *448*, 1050–1053.

García-Alcalde, F., Okonechnikov, K., Carbonell, J., Cruz, L.M., Götz, S., Tarazona, S., Dopazo, J., Meyer, T.F., and Conesa, A. (2012). Qualimap: evaluating next-generation sequencing alignment data. Bioinformatics *28*, 2678–2679.

Gel, B., and Serra, E. (2017). KaryoploteR: An R/Bioconductor package to plot customizable genomes displaying arbitrary data. Bioinformatics *33*, 3088–3090.

Gendrel, A.V., Attia, M., Chen, C.J., Diabangouaya, P., Servant, N., Barillot, E., and Heard, E. (2014). Developmental dynamics and disease potential of random monoallelic gene

expression. Dev. Cell *28*, 366–380.

Gendrel, A.V., Marion-Poll, L., Katoh, K., and Heard, E. (2016). Random monoallelic expression of genes on autosomes: Parallels with X-chromosome inactivation. Semin. Cell Dev. Biol. *56*, 100–110.

Gerull, B., Gramlich, M., Atherton, J., McNabb, M., Trombitás, K., Sasse-Klaassen, S., Seidman, J.G., Seidman, C., Granzier, H., Labeit, S., et al. (2002). Mutations of TTN, encoding the giant muscle filament titin, cause familial dilated cardiomyopathy. Nat. Genet. *30*, 201–204.

Ghaoui, R., Palmio, J., Brewer, J., Lek, M., Needham, M., Evilä, A., Hackman, P., Jonson, P.H., Penttilä, S., Vihola, A., et al. (2016). Mutations in HSPB8 causing a new phenotype of distal myopathy and motor neuropathy. Neurology *86*, 391–398.

Gibney, E.R., and Nolan, C.M. (2010). Epigenetics and gene expression. Heredity (Edinb). *105*, 4–13.

Gimelbrant, A., Hutchinson, J.N., Thompson, B.R., and Chess, A. (2007). Widespread monoallelic expression on human autosomes. Science *318*, 1136–1140.

Gimelbrant, A.A., Ensminger, A.W., Qi, P., Zucker, J., and Chess, A. (2005). Monoallelic expression and asynchronous replication of p120 catenin in mouse and human cells. J. Biol. Chem. *280*, 1354–1359.

Goldmit, M., and Bergman, Y. (2004). Monoallelic gene expression: A repertoire of recurrent themes. Immunol. Rev. *200*, 197–214.

Goodell, M.A., Brose, K., Paradis, G., Conner, A.S., and Mulligan, R.C. (1996). Isolation and functional properties of murine hematopoietic stem cells that are replicating in vivo. J. Exp. Med. *183*, 1797–1806.

Goto, Y., and Kimura, H. (2009). Inactive X chromosome-specific histone H3 modifications and CpG hypomethylation flank a chromatin boundary between an X-inactivated and an escape gene. Nucleic Acids Res. *37*, 7416–7428.

Goverman, J., Minard, K., Shastri, N., Hunkapiller, T., Hansburg, D., Sercarz, E., and Hood,

L. (1985). Rearranged β t cell receptor genes in a helper t cell clone specific for lysozyme: No correlation between Vβ and MHC restriction. Cell *40*, 859–867.

GTEx Consortium (2017). Genetic effects on gene expression across human tissues. Nature *550*, 204–213.

Gui, B., Slone, J., and Huang, T. (2017). Perspective: Is random monoallelic expression a contributor to phenotypic variability of autosomal dominant disorders? Front. Genet. *8*, 1–7.

Gupta, S., Lafontaine, D.L., Vigneau, S., Mendelevich, A., Vinogradova, S., Igarashi, K.J., Bortvin, A., Alves-Pereira, C.F., Nag, A., and Gimelbrant, A.A. (2021). RNA sequencing-based screen for reactivation of silenced alleles of autosomal genes. G3 Genes|Genomes|Genetics *jkab428*.

Haas, S., Trumpp, A., and Milsom, M.D. (2018). Causes and Consequences of Hematopoietic Stem Cell Heterogeneity. Cell Stem Cell *22*, 627–638.

Haig, D., and Westoby, M. (1989). Parent-specific gene expression and the triploid endosperm. Am. Nat. *134*, 147–155.

Hayashi, Y., Call, M.K., Liu, C.Y., Hayashi, M., Babcock, G., Ohashi, Y., and Kao, W.W.-Y. (2010). Monoallelic expression Of Krt12 gene during corneal-type epithelium differentiation Of limbal stem cells. Investig. Ophthalmol. Vis. Sci. *51*, 4562–4568.

Held, W., Roland, J., and Raulet, D.H. (1995). Allelic exclusion of Ly49-family genes encoding class I MHC-specific receptors on NK cells. Nature *376*, 355–358.

Helmrich, A., Stout-Weider, K., Hermann, K., Schrock, E., and Heiden, T. (2006). Common fragile sites are conserved features of human and mouse chromosomes and relate to large active genes. Genome Res. *16*, 1222–1230.

Holländer, G.A., Zuklys, S., Morel, C., Mizoguchi, E., Simpson, S., Terhorst, C., Wishart, W., Golan, D.E., Bhan, A.K., Mizoguchi, E., et al. (1998). Monoallelic Expression of the Interleukin-2 Locus. Science *279*, 2118–2121.

Hozumi, N., and Tonegawa, S. (1976). Evidence for somatic rearrangement of

immunoglobulin genes coding for variable and constant regions. Proc. Natl. Acad. Sci. *73*, 3628–3632.

Huang, H.C., Niu, Y., and Qin, L.X. (2015). Differential expression analysis for RNA-Seq: An overview of statistical methods and computational software. Cancer Inform. *14*, 57–67.

Jeffries, A.R., Perfect, L.W., Ledderose, J., Schalkwyk, L.C., Bray, N.J., Mill, J., and Price, J. (2012). Stochastic choice of allelic expression in human neural stem cells. Stem Cells *30*, 1938–1947.

Jeffries, A.R., Collier, D.A., Vassos, E., Curran, S., Ogilvie, C.M., and Price, J. (2013). Random or stochastic monoallelic expressed genes are enriched for neurodevelopmental disorder candidate genes. PLoS One *8*, e85093.

Jeffries, A.R., Uwanogho, D.A., Cocks, G., Perfect, L.W., Dempster, E., Mill, J., and Price, J. (2016). Erasure and reestablishment of random allelic expression imbalance after epigenetic reprogramming. RNA *22*, 1620–1630.

John, C.R., Watson, D., Russ, D., Goldmann, K., Ehrenstein, M., Pitzalis, C., Lewis, M., and Barnes, M. (2020). M3C: Monte Carlo reference-based consensus clustering. Sci. Rep. *10*, 1–14.

Johnston, P.G., and Cattanach, B.M. (1981). Controlling elements in the mouse: IV. Evidence of non-random X-inactivation. Genet. Res. *37*, 151–160.

Kambere, M.B., and Lane, R.P. (2009). Exceptional LINE density at V1R loci: The lyon repeat hypothesis revisited on autosomes. J. Mol. Evol. *68*, 145–159.

Kaneko, R., Kato, H., Kawamura, Y., Esumi, S., Hirayama, T., Hirabayashi, T., and Yagi, T. (2006). Allelic gene regulation of Pcdh-α and Pcdh-γ clusters involving both monoallelic and biallelic expression in single Purkinje cells. J. Biol. Chem. *281*, 30551–30560.

Karamitros, D., Stoilova, B., Aboukhalil, Z., Hamey, F., Reinisch, A., Samitsch, M., Quek, L., Otto, G., Repapi, E., Doondeea, J., et al. (2018). Single-cell analysis reveals the continuum of human lympho-myeloid progenitor cells article. Nat. Immunol. *19*, 85–97.

Kelly, B.L., and Locksley, R.M. (2000). Coordinate Regulation of the IL-4, IL-13, and IL-5 Cytokine Cluster in Th2 Clones Revealed by Allelic Expression Patterns. J. Immunol. *165*, 2982–2986.

Kharchenko, P. V., Silberstein, L., and Scadden, D.T. (2014). Bayesian approach to single-cell differential expression analysis. Nat. Methods *11*, 740–742.

Kiel, M.J., Yilmaz, Ö.H., Iwashita, T., Yilmaz, O.H., Terhorst, C., and Morrison, S.J. (2005). SLAM family receptors distinguish hematopoietic stem and progenitor cells and reveal endothelial niches for stem cells. Cell *121*, 1109–1121.

Kiel, M.J., Radice, G.L., and Morrison, S.J. (2007). Lack of Evidence that Hematopoietic Stem Cells Depend on N-Cadherin-Mediated Adhesion to Osteoblasts for Their Maintenance. Cell Stem Cell *1*, 204–217.

Kim, J.K., Kolodziejczyk, A.A., Illicic, T., Teichmann, S.A., and Marioni, J.C. (2015). Characterizing noise structure in single-cell RNA-seq distinguishes genuine from technical stochastic allelic expression. Nat. Commun. *6*, 1–9.

Kitamura, D., and Rajewsky, K. (1992). Targeted disruption of μ chain membrane exon causes loss of heavy-chain allelic exclusion. Nature *356*, 154–156.

Koch, C.M., Chiu, S.F., Akbarpour, M., Bharat, A., Ridge, K.M., Bartom, E.T., and Winter, D.R. (2018). A beginner's guide to analysis of RNA sequencing data. Am. J. Respir. Cell Mol. Biol. *59*, 145–157.

Kondo, M., Weissman, I.L., and Akashi, K. (1997). Identification of clonogenic common lymphoid progenitors in mouse bone marrow. Cell *91*, 661–672.

Kosak, S.T., Skok, J.A., Medina, K.L., Riblet, R., Le Beau, M.M., Fisher, A.G., and Singh, H. (2002). Subnuclear compartmentalization of immunoglobulin loci during lymphocyte development. Science *296*, 158–162.

Kratz, A., and Carninci, P. (2014). The devil in the details of RNA-seq. Nat. Biotechnol. *32*, 882–884.

Ku, C.J., Lim, K.C., Kalantry, S., Maillard, I., Engel, J.D., and Hosoya, T. (2015). A

monoallelic-to-biallelic T-cell transcriptional switch regulates GATA3 abundance. Genes Dev. *29*, 1930–1941.

Kumar, S., Deffenbacher, K., Cremers, C.W.R.J., Van Camp, G., and Kimberling, W.J. (1998). Branchio-oto-renal syndrome: Identification of novel mutations, molecular characterization, mutation distribution, and prospects for genetic testing. Genet. Test. *1*, 243–251.

Van Laer, L., Huizing, E.H., Verstreken, M., Van Zuijlen, D., Wauters, J.G., Bossuyt, P.J., Van Heyning, P. De, McGuirt, W.T., Smith, R.J.H., Willems, P.J., et al. (1998). Nonsyndromic hearing impairment is associated with a mutation in DFNA5. Nat. Genet. *20*, 194–197.

Lai, C.S.L., Fisher, S.E., Hurst, J.A., Vargha-Khadem, F., and Monaco, A.P. (2001). A forkhead-domain gene is mutated in a severe speech and language disorder. Nature *413*, 519–523.

Larsson, A.J.M., Ziegenhain, C., Hagemann-Jensen, M., Reinius, B., Jacob, T., Dalessandri, T., Hendriks, G.J., Kasper, M., and Sandberg, R. (2021). Transcriptional bursts explain autosomal random monoallelic expression and affect allelic imbalance. PLoS Comput. Biol. *17*, e1008772.

Leong, H.S., Chen, K., Hu, Y., Lee, S., Corbin, J., Pakusch, M., Murphy, J.M., Majewski, I.J., Smyth, G.K., Alexander, W.S., et al. (2013). Epigenetic regulator SMCHD1 functions as a tumor suppressor. Cancer Res. *73*, 1591–1599.

Leung, K.N., and Panning, B. (2014). X-inactivation: Xist RNA uses chromosome contacts to coat the X. Curr. Biol. *24*, R80–R82.

Li, D. (2019). Statistical Methods for RNA Sequencing Data Analysis. In Computational Biology, (Codon Publications), pp. 85–99.

Li, B., and Dewey, C.N. (2011). RSEM: Accurate transcript quantification from RNA-Seq data with or without a reference genome. BMC Bioinformatics *12*, 1–16.

Li, J., and Tibshirani, R. (2013). Finding consistent patterns: A nonparametric approach

for identifying differential expression in RNA-Seq data. Stat. Methods Med. Res. *22*, 519–536.

Li, B., Huang, Q., and Wei, G.H. (2019). The role of hox transcription factors in cancer predisposition and progression. Cancers *11*, 528.

Li, S., Labaj, P.P., Zumbo, P., Sykacek, P., Shi, W., Shi, L., Phan, J., Wu, P.Y., Wang, M., Wang, C., et al. (2014). Detecting and correcting systematic variation in large-scale RNA sequencing data. Nat. Biotechnol. *32*, 888–895.

Li, S.M., Valo, Z., Wang, J., Gao, H., Bowers, C.W., and Singer-Sam, J. (2012). Transcriptome-wide survey of mouse CNS-Derived cells reveals monoallelic expression within novel gene families. PLoS One *7*, 31751.

Loda, A., Brandsma, J.H., Vassilev, I., Servant, N., Loos, F., Amirnasr, A., Splinter, E., Barillot, E., Poot, R.A., Heard, E., et al. (2017). Genetic and epigenetic features direct differential efficiency of Xist-mediated silencing at X-chromosomal and autosomal locations. Nat. Commun. *8*, 1–16.

Lyon, M.F. (1961). Gene action in the X-chromosome of the mouse (mus musculus L.). Nature *190*, 372–373.

Lyon, M.F. (1998). X-Chromosome inactivation: A repeat hypothesis. Cytogenet. Cell Genet. *80*, 133–137.

Ma, K., Qiu, L., Mrasek, K., Zhang, J., Liehr, T., Quintana, L.G., and Li, Z. (2012). Common fragile sites: Genomic hotspots of DNA damage and carcinogenesis. Int. J. Mol. Sci. *13*, 11974–11999.

Van Der Maaten, L., and Hinton, G. (2008). Visualizing data using t-SNE. J. Mach. Learn. Res. *9*, 2579–2605.

Maduro, C., de Hoon, B., and Gribnau, J. (2016). Fitting the Puzzle Pieces: The Bigger Picture of XCI. Trends Biochem. Sci. *41*, 138–147.

Magklara, A., Yen, A., Colquitt, B.M., Clowney, E.J., Allen, W., Markenscoff-Papadimitriou, E., Evans, Z.A., Kheradpour, P., Mountoufaris, G., Carey, C., et al. (2011).

An epigenetic signature for monoallelic olfactory receptor expression. Cell *145*, 555–570.

Manz, M.G., Miyamoto, T., Akashi, K., and Weissman, I.L. (2002). Prospective isolation of human clonogenic common myeloid progenitors. Proc. Natl. Acad. Sci. U. S. A. *99*, 11872–11877.

Marinov, G.K., Williams, B.A., McCue, K., Schroth, G.P., Gertz, J., Myers, R.M., and Wold, B.J. (2014). From single-cell to cell-pool transcriptomes: Stochasticity in gene expression and RNA splicing. Genome Res. *24*, 496–510.

Marion-Poll, L., Forêt, B., Zielinski, D., Massip, F., Attia, M., Carter, A.C., Syx, L., Chang, H.Y., Gendrel, A.-V., and Heard, E. (2021). Locus specific epigenetic modalities of random allelic expression imbalance. Nat. Commun. *12*, 5330.

Mayle, A., Luo, M., Jeong, M., and Goodell, M.A. (2013). Flow cytometry analysis of murine hematopoietic stem cells. Cytom. Part A *83*, 27–37.

McGrath, J., and Solter, D. (1984). Completion of mouse embryogenesis requires both the maternal and paternal genomes. Cell *37*, 179–183.

Melchionda, S., Ahituv, N., Bisceglia, L., Sobe, T., Glaser, F., Rabionet, R., Arbones, M.L., Notarangelo, A., Di Iorio, E., Carella, M., et al. (2001). MYO6, the human homologue of the gene responsible for deafness in Snell's waltzer mice, is mutated in autosomal dominant nonsyndromic hearing loss. Am. J. Hum. Genet. *69*, 635–640.

Mendelevich, A., Vinogradova, S., Gupta, S., Mironov, A.A., Sunyaev, S.R., and Gimelbrant, A.A. (2021). Replicate sequencing libraries are important for quantification of allelic imbalance. Nat. Commun. *12*, 3370–3382.

Migeon, B.R. (2020). X-linked diseases: susceptible females. Genet. Med. *22*, 1156–1174.

Mikkola, I., Heavey, B., Horcher, M., and Busslinger, M. (2002). Reversion of B cell commitment upon loss of Pax5 expression. Science *297*, 110–113.

Morita, Y., Ema, H., and Nakauchi, H. (2010). Heterogeneity and hierarchy within the

most primitive hematopoietic stem cell compartment. J. Exp. Med. *207*, 1173–1182.

Morrison, S.J., Wandycz, A.M., Hemmati, H.D., Wright, D.E., and Weissman, I.L. (1997). Identification of a lineage of multipotent hematopoietic progenitors. Development *124*, 1929–1939.

Mortazavi, A., Williams, B.A., McCue, K., Schaeffer, L., and Wold, B. (2008). Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nat. Methods *5*, 621–628.

Mostoslavsky, R., Singh, N., Tenzen, T., Goldmit, M., Gabay, C., Elizur, S., Qi, P., Reubinoff, B.E., Chess, A., Cedar, H., et al. (2001). Asynchronous replication and allelic exclusion in the immune system. Nature *414*, 221–225.

Mould, A.W., Pang, Z., Pakusch, M., Tonks, I.D., Stark, M., Carrie, D., Mukhopadhyay, P., Seidel, A., Ellis, J.J., Deakin, J., et al. (2013). Smchd1 regulates a subset of autosomal genes subject to monoallelic expression in addition to being critical for X inactivation. Epigenetics and Chromatin *6*, 1–16.

Mousavi, M.J., Mahmoudi, M., and Ghotloo, S. (2020). Escape from X chromosome inactivation and female bias of autoimmune diseases. Mol. Med. *26*, 1–20.

Mullan, M., Crawford, F., Axelman, K., Houlden, H., Lilius, L., Winblad, B., and Lannfelt, L. (1992). A pathogenic mutation for probable Alzheimer's disease in the APP gene at the N–terminus of β–amyloid. Nat. Genet. *1*, 345–347.

Muller-Sieburg, C.E., Cho, R.H., Thoman, M., Adkins, B., and Sieburg, H.B. (2002). Deterministic regulation of hematopoietic stem cell self-renewal and differentiation. Blood *100*, 1302–1309.

Muller-Sieburg, C.E., Cho, R.H., Karlsson, L., Huang, J.F., and Sieburg, H.B. (2004). Myeloid-biased hematopoietic stem cells have extensive self-renewal capacity but generate diminished lymphoid progeny with impaired IL-7 responsiveness. Blood *103*, 4111–4118.

Nag, A., Savova, V., Fung, H.L., Miron, A., Yuan, G.C., Zhang, K., and Gimelbrant, A.A. (2013). Chromatin signature of widespread monoallelic expression. Elife e01256.

Nag, A., Vigneau, S., Savova, V., Zwemer, L.M., and Gimelbrant, A.A. (2015). Chromatin Signature Identifies Monoallelic Gene Expression Across Mammalian Cell Types. G3; Genes|Genomes|Genetics *5*, 1713–1720.

Nagai, M.H., Armelin-Correa, L.M., and Malnic, B. (2016). Monogenic and monoallelic expression of odorant receptors. Mol. Pharmacol. *90*, 633–639.

Ng, K.K.H., Yui, M.A., Mehta, A., Siu, S., Irwin, B., Pease, S., Hirose, S., Elowitz, M.B., Rothenberg, E. V., and Kueh, H.Y. (2018). A stochastic epigenetic switch controls the dynamics of T-cell lineage commitment. Elife *7*, e37851.

Nicholls, R.D., Saitoh, S., and Horsthemke, B. (1998). Imprinting in Prader-Willi and Angelman syndromes. Trends Genet. *14*, 194–200.

Nutt, S.L., Vambrie, S., Steinlein, P., Kozmik, Z., Rolink, A., Weith, A., and Busslinger, M. (1999a). Independent regulation of the two Pax5 alleles during B-cell development. Nat. Genet. *21*, 390–395.

Nutt, S.L., Heavey, B., Rolink, A.G., and Busslinger, M. (1999b). Commitment to the B-lymphoid lineage depends on the transcription factor Pax5. J. Immunol. *401*, 556–562.

Okada, S., Nakauchi, H., Nagayoshi, K., Nishikawa, S.I., Miura, Y., and Suda, T. (1992). In vivo and in vitro stem cell function of c-kit- and Sca-1-positive murine hematopoietic cells. Blood *80*, 3044–3050.

Ormundo, L.F., Machado, C.F., Sakamoto, E.D., Simões, V., and Armelin-Correa, L. (2020). LINE-1 specific nuclear organization in mice olfactory sensory neurons. Mol. Cell. Neurosci. *105*, 103494.

Osawa, M., Hanada, K.I., Hamada, H., and Nakauchi, H. (1996). Long-term lymphohematopoietic reconstitution by a single CD34- low/negative hematopoietic stem cell. Science *273*, 242–245.

Pamcer, Z., Amemiya, C.T., Ehrhardt, G.R.A., Coitlin, J., Gartland, G.L., and Cooper, M.D. (2004). Somatic diversification of variable lymphocyte receptors in the agnathan sea lamprey. Nature *430*, 174–180.

Patro, R., Duggal, G., Love, M.I., Irizarry, R.A., and Kingsford, C. (2017). Salmon provides fast and bias-aware quantification of transcript expression. Nat. Methods *14*, 417–419.

Paznekas, W.A., Karczeski, B., Vermeer, S., Lowry, R.B., Delatycki, M., Laurence, F., Koivisto, P.A., Van Maldergem, L., Boyadjiev, S.A., Bodurtha, J.N., et al. (2009). GJA1 mutations, variants, and connexin 43 dysfunction as it relates to the oculodentodigital dysplasia phenotype. Hum. Mutat. *30*, 724–733.

Pereira, J.P., Girard, R., Chaby, R., Cumano, A., and Vieira, P. (2003). Monoallelic expression of the murine gene encoding Toll-like receptor 4. Nat. Immunol. *4*, 464–470.

Pernis, B., Chiappino, G., Kelus, A.S., and Gell, P.G. (1965). Cellular localization of immunoglobulins with different allotypic specificities in rabbit lymphoid tissues. J. Exp. Med. *122*, 853–876.

Peterson, H., Kolberg, L., Raudvere, U., Kuzmin, I., and Vilo, J. (2020). gprofiler2 -- an R package for gene list functional enrichment analysis and namespace conversion toolset g: Profiler. F1000Research *9*, 709.

Pinter, S.F., Colognori, D., Beliveau, B.J., Sadreyev, R.I., Payer, B., Yildirim, E., Wu, C.T., and Lee, J.T. (2015). Allelic imbalance is a prevalent and tissue-specific feature of the mouse transcriptome. Genetics *200*, 537–549.

Platt, E.J., Smith, L., and Thayer, M.J. (2018). L1 retrotransposon antisense RNA within ASAR lncRNAs controls chromosome-wide replication timing. J. Cell Biol. *217*, 541–553.

Polymeropoulos, M.H., Lavedan, C., Leroy, E., Ide, S.E., Dehejia, A., Dutra, A., Pike, B., Root, H., Rubenstein, J., Boyer, R., et al. (1997). Mutation in the α-synuclein gene identified in families with Parkinson's disease. Science *276*, 2045–2047.

Raudvere, U., Kolberg, L., Kuzmin, I., Arak, T., Adler, P., Peterson, H., and Vilo, J. (2019). G:Profiler: A web server for functional enrichment analysis and conversions of gene lists (2019 update). Nucleic Acids Res. *47*, W191–W198.

Reik, W., and Walter, J. (2001). Genomic imprinting: Parental influence on the genome. Nat. Rev. Genet. *2*, 21–32.

Reinius, B., and Sandberg, R. (2015). Random monoallelic expression of autosomal genes: Stochastic transcription and allele-level regulation. Nat. Rev. Genet. *16*, 653–664.

Reinius, B., and Sandberg, R. (2018). Reply to 'High prevalence of clonal monoallelic expression.' Nat. Genet. *50*, 1199–1200.

Reinius, B., Mold, J.E., Ramsköld, D., Deng, Q., Johnsson, P., Michaëlsson, J., Frisén, J., and Sandberg, R. (2016). Analysis of allelic expression patterns in clonal somatic cells by single-cell RNA-seq. Nat. Genet. *48*, 1430–1435.

Renfree, M.B., Hore, T.A., Shaw, G., Marshall Graves, J.A., and Pask, A.J. (2009). Evolution of genomic imprinting: Insights from marsupials and monotremes. Annu. Rev. Genomics Hum. Genet. *10*, 241–262.

Reya, T. (2003). Regulation of Hematopoietic Stem Cell Self-Renewal. Recent Prog. Horm. Res. *58*, 283–295.

Robinson, M.D., and Oshlack, A. (2010). A scaling normalization method for differential expression analysis of RNA-seq data. Genome Biol. *11*, 1–9.

Rodriguez-Fraticelli, A.E., Wolock, S.L., Weinreb, C.S., Panero, R., Patel, S.H., Jankovic, M., Sun, J., Calogero, R.A., Klein, A.M., and Camargo, F.D. (2018). Clonal analysis of lineage fate in native haematopoiesis. Nature *553*, 212–216.

Rodriguez, I., Feinstein, P., and Mombaerts, P. (1999). Variable patterns of axonal projections of sensory neurons in the mouse vomeronasal system. Cell *97*, 199–208.

Rolink, A.G., Nutt, S.L., Melchers, F., and Busslinger, M. (1999). Long-term in vivo reconstitution of T-cell development by Pax5-deficient B-cell progenitors. Nature *401*, 603–606.

Rovelet-Lecrux, A., Hannequin, D., Raux, G., Le Meur, N., Laquerrière, A., Vital, A., Dumanchin, C., Feuillette, S., Brice, A., Vercelletto, M., et al. (2006). APP locus duplication causes autosomal dominant early-onset Alzheimer disease with cerebral amyloid angiopathy. Nat. Genet. *38*, 24–26.

Ruf, R.G., Xu, P.X., Silvius, D., Otto, E.A., Beekmann, F., Muerb, U.T., Kumar, S., Neuhaus,

T.J., Kemper, M.J., Raymond, R.M., et al. (2004). SIX1 mutations cause branchio-oto-renal syndrome by disruption of EYA1-SIX1-DNA complexes. Proc. Natl. Acad. Sci. U. S. A. *101*, 8090–8095.

Rv, P., Sundaresh, A., Karunyaa, M., Arun, A., and Gayen, S. (2021). Autosomal Clonal Monoallelic Expression: Natural or Artifactual? Trends Genet. *37*, 206–211.

Sado, T., Okano, M., Li, E., and Sasaki, H. (2004). De novo DNA methylation is dispensable for the initiation and propagation of X chromosome inactivation. Development *131*, 975–982.

Saleh, A., Makrigiannis, A.P., Hodge, D.L., and Anderson, S.K. (2002). Identification of a Novel Ly49 Promoter That Is Active in Bone Marrow and Fetal Thymus. J. Immunol. *168*, 5163–5169.

Saleh, A., Davies, G.E., Pascal, V., Wright, P.W., Hodge, D.L., Cho, E.H., Lockett, S.J., Abshari, M., and Anderson, S.K. (2004). Identification of probabilistic transcriptional switches in the Ly49 gene cluster: A eukaryotic mechanism for selective gene activation. Immunity *21*, 55–66.

Sano, Y., Shimada, T., Nakashima, H., Nicholson, R.H., Eliason, J.F., Kocarek, T.A., and Ko, M.S.H. (2001). Random monoallelic expression of three genes clustered within 60 kb of mouse t complex genomic DNA. Genome Res. *11*, 1833–1841.

Savarese, F., Flahndorfer, K., Jaenisch, R., Busslinger, M., and Wutz, A. (2006). Hematopoietic Precursor Cells Transiently Reestablish Permissiveness for XInactivation. Mol. Cell. Biol. *26*, 7167–7177.

Savol, A.J., Wang, P.I., Jeon, Y., Colognori, D., Yildirim, E., Pinter, S.F., Payer, B., Lee, J.T., and Sadreyev, R.I. (2017). Genome-wide identification of autosomal genes with allelic imbalance of chromatin state. PLoS One *12*, e0182568.

Savova, V., Chun, S., Sohail, M., Mccole, R.B., Witwicki, R., Gai, L., Lenz, T.L., Wu, C.T., Sunyaev, S.R., and Gimelbrant, A.A. (2016). Genes with monoallelic expression contribute disproportionately to genetic diversity in humans. Nat. Genet. *48*, 231–237.

Schaniel, C., Bruno, L., Melchers, F., and Rolink, A.G. (2002). Multiple hematopoietic cell lineages develop in vivo from transplanted Pax5-deficient pre-B I-cell clones. Blood *99*, 472–478.

Schlesinger, S., Selig, S., Bergman, Y., and Cedar, H. (2009). Allelic inactivation of rDNA loci. Genes Dev. *23*, 2437–2447.

Schroeder, T. (2010). Hematopoietic Stem Cell Heterogeneity: Subtypes, Not Unpredictable Behavior. Cell Stem Cell *6*, 203–207.

Sebat, J., Lakshmi, B., Malhotra, D., Troge, J., Lese-Martin, C., Walsh, T., Yamrom, B., Yoon, S., Krasnitz, A., Kendall, J., et al. (2007). Strong association of de novo copy number mutations with autism. Science *316*, 445–449.

Selcen, D., Muntoni, F., Burton, B.K., Pegoraro, E., Sewry, C., Bite, A. V., and Engel, A.G. (2009). Mutation in BAG3 causes severe dominant childhood muscular dystrophy. Ann. Neurol. *65*, 83–89.

Serizawa, S., Miyamichi, K., Nakatani, H., Suzuki, M., Saito, M., Yoshihara, Y., and Sakano, H. (2003). Negative Feedback Regulation Ensures the One Receptor-One Olfactory Neuron Rule in Mouse. Science *302*, 2088–2094.

Sieburg, H.B., Cho, R.H., Dykstra, B., Uchida, N., Eaves, C.J., and Muller-Sieburg, C.E. (2006). The hematopoietic stem compartment consists of a limited number of discrete stem cell subsets. Blood *107*, 2311–2316.

Sierra, I., and Anguera, M.C. (2019). Enjoy the silence: X-chromosome inactivation diversity in somatic cells. Curr. Opin. Genet. Dev. *55*, 26–31.

Simon, M.D., Pinter, S.F., Fang, R., Sarma, K., Rutenberg-Schoenberg, M., Bowman, S.K., Kesner, B.A., Maier, V.K., Kingston, R.E., and Lee, J.T. (2013). High-resolution Xist binding maps reveal two-step spreading during X-chromosome inactivation. Nature *504*, 465–469.

Singh, U., and Wurtele, E.S. (2021). orfipy: a fast and flexible tool for extracting ORFs. Bioinformatics *37*, 3019–3020.

Singh, N., Ebrahimi, F.A.W., Gimelbrant, A.A., Ensminger, A.W., Tackett, M.R., Qi, P., Gribnau, J., and Chess, A. (2003). Coordination of the random asynchronous replication of autosomal loci. Nat. Genet. *33*, 339–341.

Singleton, A.B., Farrer, M., Johnson, J., Singleton, A., Hague, S., Kachergus, J., Hulihan, M., Peuralinna, T., Dutra, A., Nussbaum, R., et al. (2003). α-Synuclein Locus Triplication Causes Parkinson's Disease. Science *302*, 841.

Skok, J.A., Brown, K.E., Azuara, V., Caparros, M.L., Baxter, J., Takacs, K., Dillon, N., Gray, D., Perry, R.P., Merkenschlager, M., et al. (2001). Nonequivalent nuclear location of immunoglobulin alleles in B lymphocytes. Nat. Immunol. *2*, 848–854.

Souyris, M., Cenac, C., Azar, P., Daviaud, D., Canivet, A., Grunenwald, S., Pienkowski, C., Chaumeil, J., Mejía, J.E., and Guéry, J.C. (2018). TLR7 escapes X chromosome inactivation in immune cells. Sci. Immunol. *3*, eaap8855.

Splinter, E., de Wit, E., Nora, E.P., Klous, P., van de Werken, H.J.G., Zhu, Y., Kaaij, L.J.T., van Ijcken, W., Gribnau, J., Heard, E., et al. (2011). The inactive X chromosome adopts a unique three-dimensional conformation that is dependent on Xist RNA. Genes Dev. *25*, 1371–1383.

Stoffregen, E.P., Donley, N., Stauffer, D., Smith, L., and Thayer, M.J. (2011). An autosomal locus that controls chromosomewide replication timing and mono-allelic expression. Hum. Mol. Genet. *20*, 2366–2378.

Surani, M.A.H., Barton, S.C., and Norris, M.L. (1984). Development of reconstituted mouse eggs suggests imprinting of the genome during gametogenesis. Nature *308*, 548–550.

Syrett, C.M., Sindhava, V., Hodawadekar, S., Myles, A., Liang, G., Zhang, Y., Nandi, S., Cancro, M., Atchison, M., and Anguera, M.C. (2017). Loss of Xist RNA from the inactive X during B cell development is restored in a dynamic YY1-dependent two-step process in activated B cells. PLoS Genet. *13*.

Syrett, C.M., Paneru, B., Sandoval-Heglund, D., Wang, J., Banerjee, S., Sindhava, V., Behrens, E.M., Atchison, M., and Anguera, M.C. (2019). Altered X-chromosome

inactivation in T cells may promote sex-biased autoimmune diseases. JCI Insight *4*, e126751.

Takagi, N. (1974). Differentiation of X chromosomes in early female mouse embryos. Exp. Cell Res. *86*, 127–135.

Takizawa, T., Gudla, P.R., Guo, L., Lockett, S., and Misteli, T. (2008). Allele-specific nuclear positioning of the monoallelically expressed astrocyte marker GFAP. Genes Dev. *22*, 489–498.

Tannan, N.B., Brahmachary, M., Garg, P., Borel, C., Alnefaie, R., Watson, C.T., Simon Thomas, N., and Sharp, A.J. (2014). Dna methylation profiling in X;autosome translocations supports a role for L1 repeats in the spread of X chromosome inactivation. Hum. Mol. Genet. *23*, 1224–1236.

Tarazona, S., Furió-Tarí, P., Turrà, D., Di Pietro, A., Nueda, M.J., Ferrer, A., and Conesa, A. (2015). Data quality aware analysis of differential expression in RNA-seq with NOISeq R/Bioc package. Nucleic Acids Res. *43*, e140–e140.

Tasic, B., Nabholz, C.E., Baldwin, K.K., Kim, Y., Rueckert, E.H., Ribich, S.A., Cramer, P., Wu, Q., Axel, R., and Maniatis, T. (2002). Promoter choice determines splice site selection in protocadherin α and γ pre-mRNA splicing. Mol. Cell *10*, 21–33.

Thomas, B.J., Rubio, E.D., Krumm, N., Broin, P.Ó., Bomsztyk, K., Welcsh, P., Greally, J.M., Golden, A.A., and Krumm, A. (2011). Allele-specific transcriptional elongation regulates monoallelic expression of the IGF2BP1 gene. Epigenetics and Chromatin *4*, 14.

Till, J.E., and McCulloch, E.A. (1961). A direct measurement of the radiation sensitivity of normal mouse bone marrow cells. Radiat. Res. *14*, 213–222.

Trapnell, C., Williams, B.A., Pertea, G., Mortazavi, A., Kwan, G., Van Baren, M.J., Salzberg, S.L., Wold, B.J., and Pachter, L. (2010). Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nat. Biotechnol. *28*, 511–515.

Uchida, N., Dykstra, B., Lyons, K.J., Leung, F.Y.K., and Eaves, C.J. (2003). Different in vivo

repopulating activities of purified hematopoietic stem cells before and after being stimulated to divide in vitro with the same kinetics. Exp. Hematol. *31*, 1338–1347.

Velten, L., Haas, S.F., Raffel, S., Blaszkiewicz, S., Islam, S., Hennig, B.P., Hirche, C., Lutz, C., Buss, E.C., Nowak, D., et al. (2017). Human haematopoietic stem cell lineage commitment is a continuous process. Nat. Cell Biol. *19*, 271–281.

Verovskaya, E., Broekhuis, M.J.C., Zwart, E., Weersing, E., Ritsema, M., Bosman Lisette J., J., Van Poele, T., De Haan, G., and Bystrykh, L. V. (2014). Asymmetry in skeletal distribution of mouse hematopoietic stem cell clones and their equilibration by mobilizing cytokines. J. Exp. Med. *211*, 487–497.

Vettermann, C., and Schlissel, M.S. (2010). Allelic exclusion of immunoglobulin genes: Models and mechanisms. Immunol. Rev. *237*, 22–42.

Vigneau, S., Vinogradova, S., Savova, V., and Gimelbrant, A. (2018). High prevalence of clonal monoallelic expression. Nat. Genet. *50*, 1198–1199.

Vincent, A., Forster, N., Maynes, J.T., Paton, T.A., Billingsley, G., Roslin, N.M., Ali, A., Sutherland, J., Wright, T., Westall, C.A., et al. (2014). OTX2 mutations cause autosomal dominant pattern dystrophy of the retinal pigment epithelium. J. Med. Genet. *51*, 797–805.

Wade, C.M., Kulbokas, E.J., Kirby, A.W., Zody, M.C., Mullikin, J.C., Lander, E.S., Lindblad-Toh, K., and Daly, M.J. (2002). The mosaic structure of variation in the laboratory mouse genome. Nature *420*, 574–578.

Wagers, A.J., Sherwood, R.I., Christensen, J.L., and Weissman, I.L. (2002). Little evidence for developmental plasticity of adult hematopoietic stem cells. Science *297*, 2256–2259.

Wagner, G.P., Kin, K., and Lynch, V.J. (2012). Measurement of mRNA abundance using RNA-seq data: RPKM measure is inconsistent among samples. Theory Biosci. *131*, 281–285.

Wang, F., Nemes, A., Mendelsohn, M., and Axel, R. (1998). Odorant receptors govern the formation of a precise topographic map. Cell *93*, 47–60.

Wang, J., Valo, Z., Smith, D., and Singer-Sam, J. (2007). Monoallelic expression of multiple genes in the CNS. PLoS One *2*, e1293.

Wang, J., Syrett, C.M., Kramer, M.C., Basu, A., Atchison, M.L., and Anguera, M.C. (2016). Unusual maintenance of X chromosome inactivation predisposes female lymphocytes for increased expression from the inactive X. Proc. Natl. Acad. Sci. U. S. A. *113*, E2029–E2038.

Wang, X., Su, H., and Bradley, A. (2002). Molecular mechanisms governing Pcdh-γ gene expression: Evidence for a multiple promoter and cis-alternative splicing model. Genes Dev. *16*, 1890–1905.

Wayne, S., Robertson, N.G., DeClau, F., Chen, N., Verhoeven, K., Prasad, S., Tranebjärg, L., Morton, C.C., Ryan, A.F., Van Camp, G., et al. (2001). Mutations in the transcriptional activator EYA4 cause late-onset deafness at the DFNA 10 locus. Hum. Mol. Genet. *10*, 195–200.

West, J.D., and Chapman, V.M. (1978). Variation for X chromosome expression in mice detected by electrophoresis of phosphoglycerate kinase. Genet. Res. *32*, 91–102.

Wibom, R., Lasorsa, F.M., Töhönen, V., Barbaro, M., Sterky, F.H., Kucinski, T., Naess, K., Jonsson, M., Pierri, C.L., Palmieri, F., et al. (2009). AGC1 Deficiency Associated with Global Cerebral Hypomyelination. N. Engl. J. Med. *361*, 489–495.

Wilkinson, A.C., Igarashi, K.J., and Nakauchi, H. (2020). Haematopoietic stem cell self-renewal in vivo and ex vivo. Nat. Rev. Genet. *21*, 541–554.

Wilson, N.K., Kent, D.G., Buettner, F., Shehata, M., Macaulay, I.C., Calero-Nieto, F.J., Sánchez Castillo, M., Oedekoven, C.A., Diamanti, E., Schulte, R., et al. (2015). Combined Single-Cell Functional and Gene Expression Analysis Resolves Heterogeneity within Stem Cell Populations. Cell Stem Cell *16*, 712–724.

Wolf, J.B., and Hager, R. (2006). A maternal-offspring coadaptation theory for the evolution of genomic imprinting. PLoS Biol. *4*, 2238–2243.

Wu, H., Luo, J., Yu, H., Rattner, A., Mo, A., Wang, Y., Smallwood, P.M., Erlanger, B.,

Wheelan, S.J., and Nathans, J. (2014). Cellular Resolution Maps of X Chromosome Inactivation: Implications for Neural Development, Function, and Disease. Neuron *81*, 103–119.

Yamamoto, R., Morita, Y., Ooehara, J., Hamanaka, S., Onodera, M., Rudolph, K.L., Ema, H., and Nakauchi, H. (2013). Clonal analysis unveils self-renewing lineage-restricted progenitors generated directly from hematopoietic stem cells. Cell *154*, 1112–1126.

Yamamoto, R., Wilkinson, A.C., Ooehara, J., Lan, X., Lai, C.Y., Nakauchi, Y., Pritchard, J.K., and Nakauchi, H. (2018). Large-Scale Clonal Analysis Resolves Aging of the Mouse Hematopoietic Stem Cell Compartment. Cell Stem Cell *22*, 600-607.e4.

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

Yang, L., Bryder, D., Adolfsson, J., Nygren, J., Månsson, R., Sigvardsson, M., and Jacobsen, S.E.W. (2005). Identification of Lin-Sca1+kit+CD34 +Flt3- short-term hematopoietic stem cells capable of rapidly reconstituting and rescuing myeloablated transplant recipients. Blood *105*, 2717–2723.

Yokota, T. (2019). “Hierarchy” and “Holacracy”; A Paradigm of the Hematopoietic System. Cells *8*, 1138.

Yu, V.W.C., Yusuf, R.Z., Oki, T., Wu, J., Saez, B., Wang, X., Cook, C., Baryawno, N., Ziller, M.J., Lee, E., et al. (2016). Epigenetic Memory Underlies Cell-Autonomous Heterogeneous Behavior of Hematopoietic Stem Cells. Cell *167*, 1310-1322.e17.

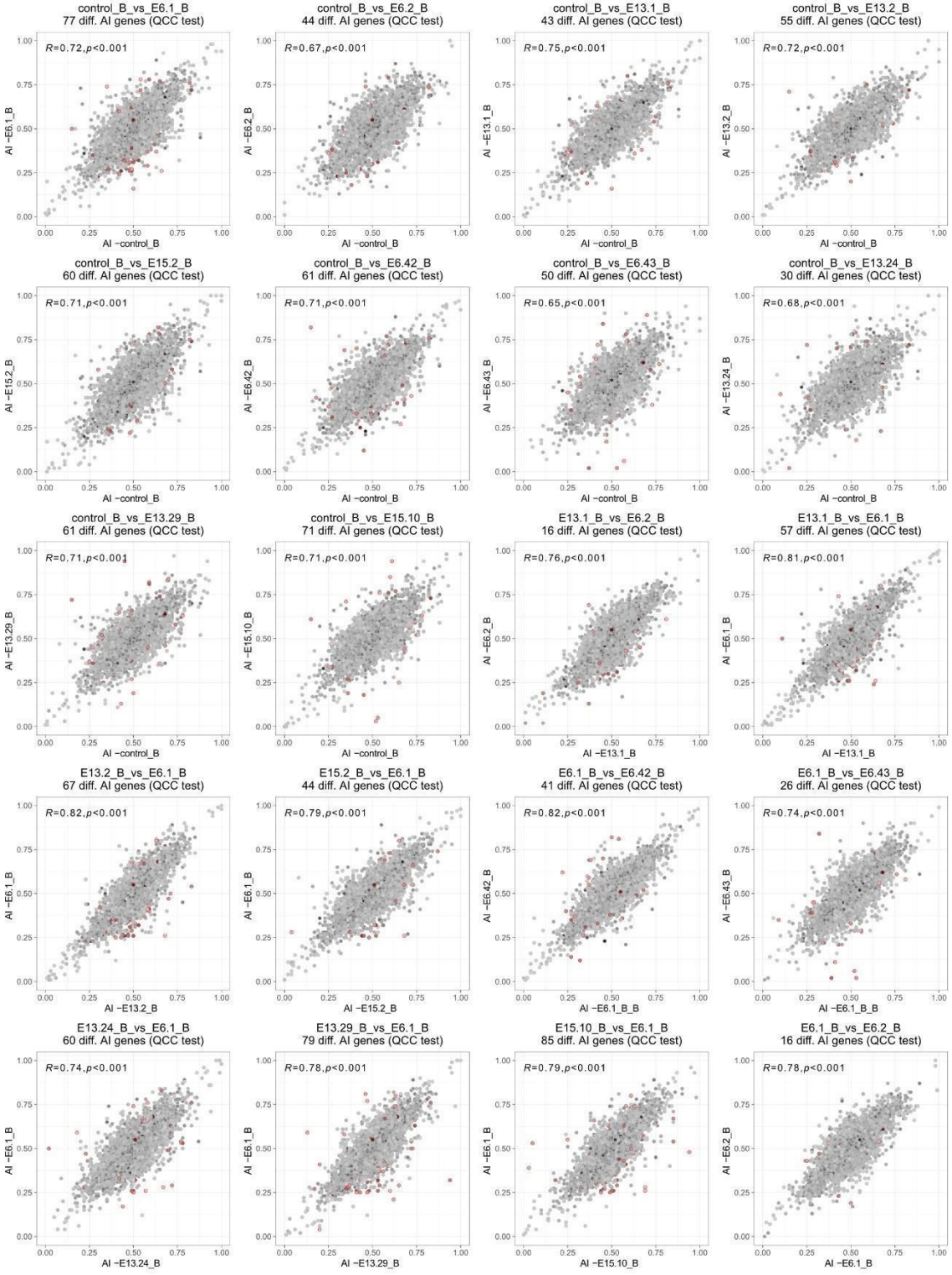
Zhang, X., and Firestein, S. (2002). The olfactory receptor gene superfamily of the mouse. Nat. Neurosci. *5*, 124–133.

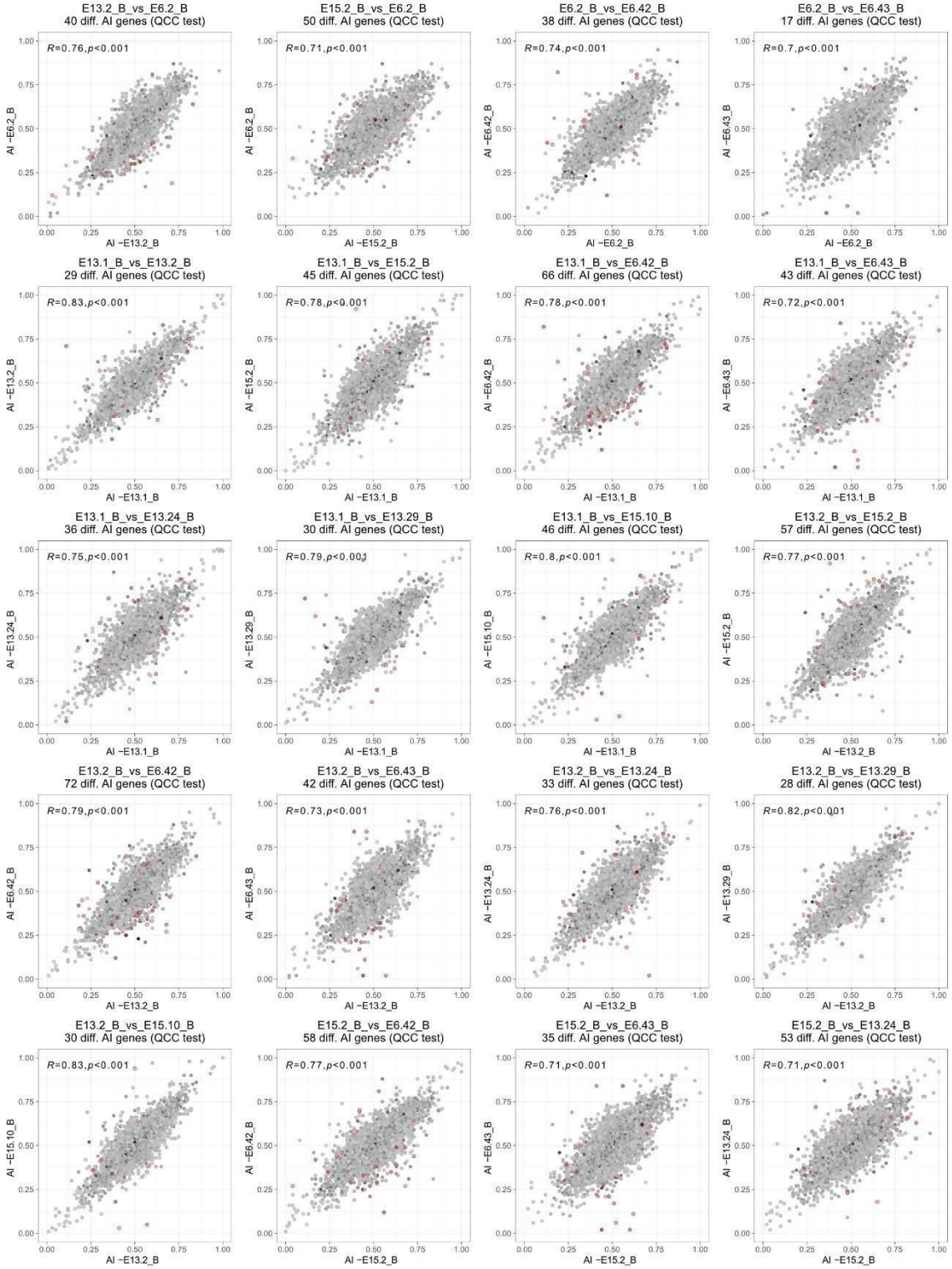
Zito, A., Davies, M.N., Tsai, P.C., Roberts, S., Andres-Ejarque, R., Nardone, S., Bell, J.T., Wong, C.C.Y., and Small, K.S. (2019). Heritability of skewed X-inactivation in female twins is tissue-specific and associated with age. Nat. Commun. *10*, 1–11.

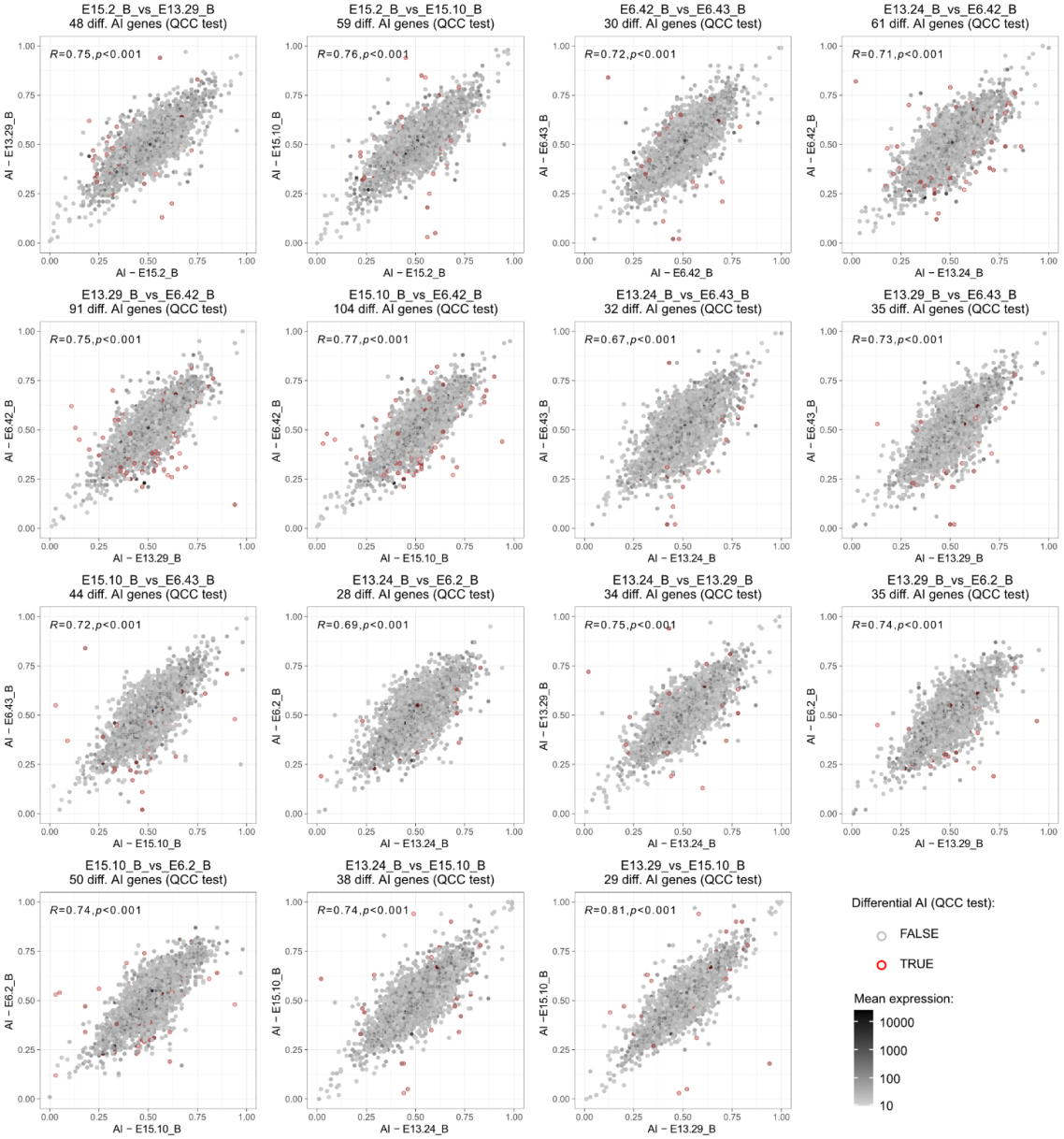
Zwemer, L.M., Zak, A., Thompson, B.R., Kirby, A., Daly, M.J., Chess, A., and Gimelbrant, A.A. (2012). Autosomal monoallelic expression in the mouse. Genome Biol. *13*, R10.

7. Supplementary results

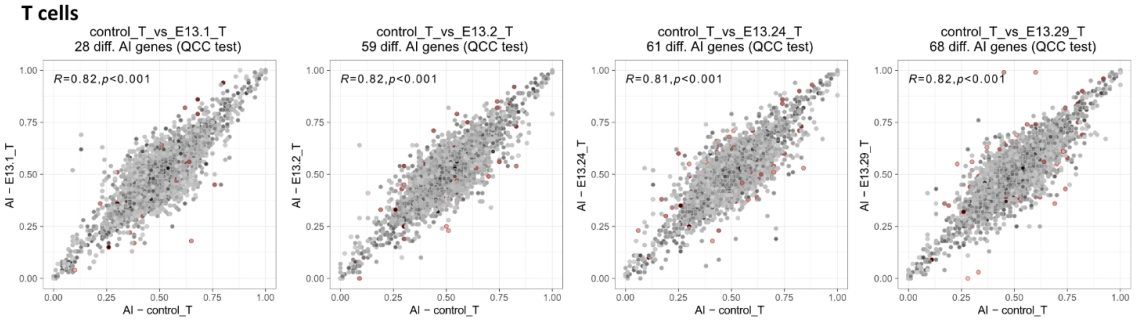
B cells

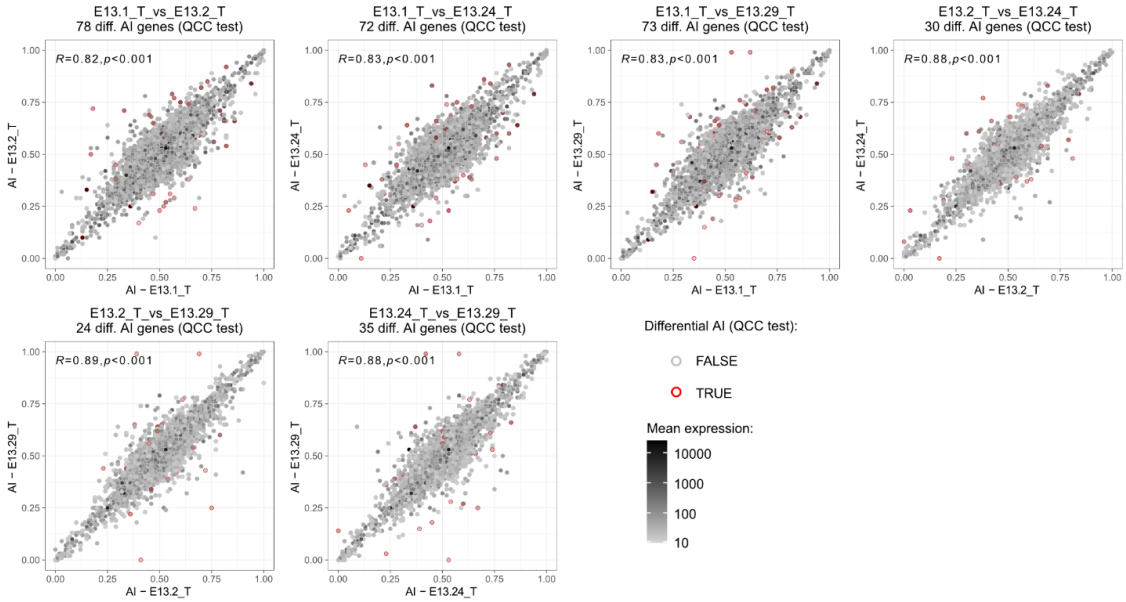

control_B_vs_E6.1_B
77 diff. AI genes (QCC test)
R=0.72,p<0.001
AI −control_B
AI −E6.1_B
control_B_vs_E6.2_B
44 diff. AI genes (QCC test)
R=0.67,p<0.001
AI −control_B
AI −E6.2_B
control_B_vs_E13.1_B
43 diff. AI genes (QCC test)
R=0.75,p<0.001
AI −control_B
AI −E13.1_B
control_B_vs_E13.2_B
55 diff. AI genes (QCC test)
R=0.72,p<0.001
AI −control_B
AI −E13.2_B
control_B_vs_E15.2_B
60 diff. AI genes (QCC test)
R=0.71,p<0.001
AI −control_B
AI −E15.2_B
control_B_vs_E6.42_B
61 diff. AI genes (QCC test)
R=0.71,p<0.001
AI −control_B
AI −E6.42_B
control_B_vs_E6.43_B
50 diff. AI genes (QCC test)
R=0.65,p<0.001
AI −control_B
AI −E6.43_B
control_B_vs_E13.24_B
30 diff. AI genes (QCC test)
R=0.68,p<0.001
AI −control_B
AI −E13.24_B
control_B_vs_E13.29_B
61 diff. AI genes (QCC test)
R=0.71,p<0.001
AI −control_B
AI −E13.29_B
control_B_vs_E15.10_B
71 diff. AI genes (QCC test)
R=0.71,p<0.001
AI −control_B
AI −E15.10_B
E13.1_B_vs_E6.2_B
16 diff. AI genes (QCC test)
R=0.76,p<0.001
AI −E13.1_B
AI −E6.2_B
E13.1_B_vs_E6.1_B
57 diff. AI genes (QCC test)
R=0.81,p<0.001
AI −E13.1_B
AI −E6.1_B
E13.2_B_vs_E6.1_B
67 diff. AI genes (QCC test)
R=0.82,p<0.001
AI −E13.2_B
AI −E6.1_B
E15.2_B_vs_E6.1_B
44 diff. AI genes (QCC test)
R=0.79,p<0.001
AI −E15.2_B
AI −E6.1_B
E6.1_B_vs_E6.42_B
41 diff. AI genes (QCC test)
R=0.82,p<0.001
AI −E6.1_B_B
AI −E6.42_B
E6.1_B_vs_E6.43_B
26 diff. AI genes (QCC test)
R=0.74,p<0.001
AI −E6.1_B_B
AI −E6.43_B
E13.24_B_vs_E6.1_B
60 diff. AI genes (QCC test)
R=0.74,p<0.001
AI −E13.24_B
AI −E6.1_B
E13.29_B_vs_E6.1_B
79 diff. AI genes (QCC test)
R=0.78,p<0.001
AI −E13.29_B
AI −E6.1_B
E15.10_B_vs_E6.1_B
85 diff. AI genes (QCC test)
R=0.79,p<0.001
AI −E15.10_B
AI −E6.1_B
E6.1_B_vs_E6.2_B
16 diff. AI genes (QCC test)
R=0.78,p<0.001
AI −E6.1_B
AI −E6.2_B


E13.2_B_vs_E6.2_B
40 diff. AI genes (QCC test)
R=0.76,p<0.001
AI −E6.2_B
AI −E13.2_B
E15.2_B_vs_E6.2_B
50 diff. AI genes (QCC test)
R=0.71,p<0.001
AI −E6.2_B
AI −E15.2_B
E6.2_B_vs_E6.42_B
38 diff. AI genes (QCC test)
R=0.74,p<0.001
AI −E6.42_B
AI −E6.2_B
E6.2_B_vs_E6.43_B
17 diff. AI genes (QCC test)
R=0.7,p<0.001
AI −E6.43_B
AI −E6.2_B
E13.1_B_vs_E13.2_B
29 diff. AI genes (QCC test)
R=0.83,p<0.001
AI −E13.2_B
AI −E13.1_B
E13.1_B_vs_E15.2_B
45 diff. AI genes (QCC test)
R=0.78,p<0.001
AI −E15.2_B
AI −E13.1_B
E13.1_B_vs_E6.42_B
66 diff. AI genes (QCC test)
R=0.78,p<0.001
AI −E6.42_B
AI −E13.1_B
E13.1_B_vs_E6.43_B
43 diff. AI genes (QCC test)
R=0.72,p<0.001
AI −E6.43_B
AI −E13.1_B
E13.1_B_vs_E13.24_B
36 diff. AI genes (QCC test)
R=0.75,p<0.001
AI −E13.24_B
AI −E13.1_B
E13.1_B_vs_E13.29_B
30 diff. AI genes (QCC test)
R=0.79,p<0.001
AI −E13.29_B
AI −E13.1_B
E13.1_B_vs_E15.10_B
46 diff. AI genes (QCC test)
R=0.8,p<0.001
AI −E15.10_B
AI −E13.1_B
E13.2_B_vs_E15.2_B
57 diff. AI genes (QCC test)
R=0.77,p<0.001
AI −E15.2_B
AI −E13.2_B
E13.2_B_vs_E6.42_B
72 diff. AI genes (QCC test)
R=0.79,p<0.001
AI −E6.42_B
AI −E13.2_B
E13.2_B_vs_E6.43_B
42 diff. AI genes (QCC test)
R=0.73,p<0.001
AI −E6.43_B
AI −E13.2_B
E13.2_B_vs_E13.24_B
33 diff. AI genes (QCC test)
R=0.76,p<0.001
AI −E13.24_B
AI −E13.2_B
E13.2_B_vs_E13.29_B
28 diff. AI genes (QCC test)
R=0.82,p<0.001
AI −E13.29_B
AI −E13.2_B
E13.2_B_vs_E15.10_B
30 diff. AI genes (QCC test)
R=0.83,p<0.001
AI −E15.10_B
AI −E13.2_B
E15.2_B_vs_E6.42_B
58 diff. AI genes (QCC test)
R=0.77,p<0.001
AI −E6.42_B
AI −E15.2_B
E15.2_B_vs_E6.43_B
35 diff. AI genes (QCC test)
R=0.71,p<0.001
AI −E6.43_B
AI −E15.2_B
E15.2_B_vs_E13.24_B
53 diff. AI genes (QCC test)
R=0.71,p<0.001
AI −E13.24_B
AI −E15.2_B

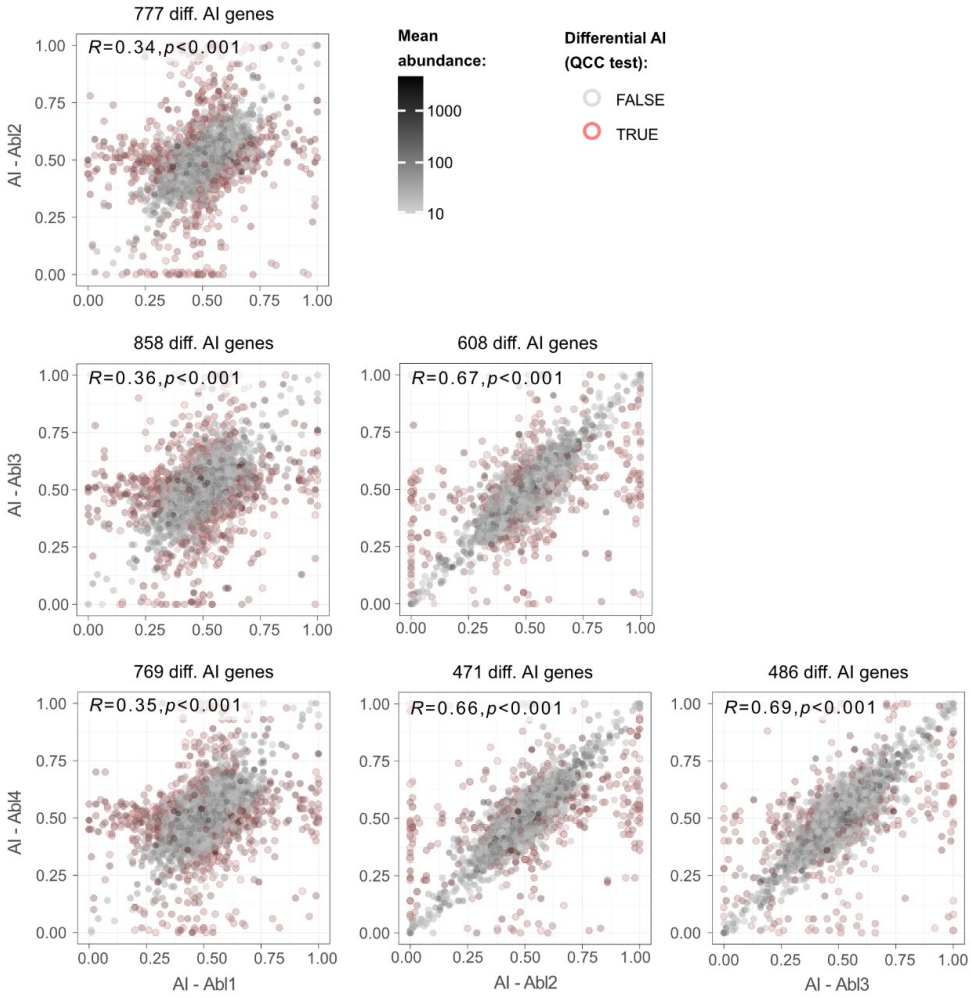

E15.2_B_vs_E13.29_B
48 diff. AI genes (QCC test)
R=0.75,p<0.001
AI – E13.29_B
AI – E15.2_B
E15.2_B_vs_E15.10_B
59 diff. AI genes (QCC test)
R=0.76,p<0.001
AI – E15.10_B
AI – E15.2_B
E6.42_B_vs_E6.43_B
30 diff. AI genes (QCC test)
R=0.72,p<0.001
AI – E6.43_B
AI – E6.42_B
E13.24_B_vs_E6.42_B
61 diff. AI genes (QCC test)
R=0.71,p<0.001
AI – E6.42_B
AI – E13.24_B
E13.29_B_vs_E6.42_B
91 diff. AI genes (QCC test)
R=0.75,p<0.001
AI – E6.42_B
AI – E13.29_B
E15.10_B_vs_E6.42_B
104 diff. AI genes (QCC test)
R=0.77,p<0.001
AI – E6.42_B
AI – E15.10_B
E13.24_B_vs_E6.43_B
32 diff. AI genes (QCC test)
R=0.67,p<0.001
AI – E6.43_B
AI – E13.24_B
E13.29_B_vs_E6.43_B
35 diff. AI genes (QCC test)
R=0.73,p<0.001
AI – E6.43_B
AI – E13.29_B
E15.10_B_vs_E6.43_B
44 diff. AI genes (QCC test)
R=0.72,p<0.001
AI – E6.43_B
AI – E15.10_B
E13.24_B_vs_E6.2_B
28 diff. AI genes (QCC test)
R=0.69,p<0.001
AI – E6.2_B
AI – E13.24_B
E13.24_B_vs_E13.29_B
34 diff. AI genes (QCC test)
R=0.75,p<0.001
AI – E13.29_B
AI – E13.24_B
E13.29_B_vs_E6.2_B
35 diff. AI genes (QCC test)
R=0.74,p<0.001
AI – E6.2_B
AI – E13.29_B
E15.10_B_vs_E6.2_B
50 diff. AI genes (QCC test)
R=0.74,p<0.001
AI – E6.2_B
AI – E15.10_B
E13.24_B_vs_E15.10_B
38 diff. AI genes (QCC test)
R=0.74,p<0.001
AI – E15.10_B
AI – E13.24_B
E13.29_vs_E15.10_B
29 diff. AI genes (QCC test)
R=0.81,p<0.001
AI –E15.10_B
AI – E13.29_B
Differential AI (QCC test):
FALSE
TRUE
Mean expression:
10000
1000
100
10

T cells


control_T_vs_E13.1_T
28 diff. AI genes (QCC test)
R=0.82,p<0.001
AI – E13.1_T
AI – control_T
control_T_vs_E13.2_T
59 diff. AI genes (QCC test)
R=0.82,p<0.001
AI – E13.2_T
AI – control_T
control_T_vs_E13.24_T
61 diff. AI genes (QCC test)
R=0.81,p<0.001
AI – E13.24_T
AI – control_T
control_T_vs_E13.29_T
68 diff. AI genes (QCC test)
R=0.82,p<0.001
AI – E13.29_T
AI – control_T



Supplementary Figure 7. 1. Pairwise comparisons of allelic imbalance between animals for B and T cells, with values of Pearson's coefficient correlation and the number of genes with a significant differential AI after applying QCC correction on the binomial tests. Abundance values are TMM-normalized counts.



Supplementary Figure 7. 2. Pairwise comparisons of allelic imbalance between Abelson-immortalized B-cell clones, with values of Pearson's coefficient correlation and the number of genes with a significant differential AI after applying QCC correction on the binomial tests. Abundance values are TMM-normalized counts.

Supplementary Table 7. 1. XCI escapees identified by other studies and our study showing values of allelic imbalance and expression in HSC-derived lymphocytes in vivo from our NGS data. Genes in bold are escapees identified in our study. Genes in red were not expressed in our samples and therefore were excluded from our study. Genes in blue were not included in our study due to lack of SNPs to estimate allelic imbalance. Read abundance corresponds to mean abundance in TMM-normalized counts.

	Allelic imbalance								
	T cells		B cells						
gene	**E13.2**	**E13.29**	**E13.24**	**E13.29**	**E15.10**	**E6.42**	**E6.43**	**Mean abundance**	**Xi expression**
1810030O07Rik	1.00	1.00	0.98	0.96	0.98	0.01	0.03	62.94	
5530601H04Rik	**0.73**	**0.74**	**0.73**	**0.74**	**0.78**	**0.16**	**0.17**	**29.65**	**yes**
5730416F02Rik	-	-	-	-	-	-	-	0.23	
Abcb7	0.99	1.00	0.99	0.97	0.99	0.00	0.01	36.13	
Alg13	0.95	0.97	0.81	0.93	0.99	0.09	0.10	14.65	
Amot	0.42	-	0.77	-	-	0.32	0.28	1.93	
Ap1s2	0.97	1.00	0.93	0.90	0.96	0.01	0.01	30.45	
Atp7a	0.95	0.99	0.82	0.93	0.99	0.04	0.10	47.07	
AU015836	-	-	-	-	-	-	-	0.49	
AU022751	-	-	-	0.00	-	-	-	0.21	
BC022960	0.98	0.94	1.00	0.89	0.96	-	-	0.86	
Bcor	1.00	1.00	0.93	0.95	0.97	0.01	0.01	71.23	
Bgn	-	-	-	-	-	-	-	0.19	
Bhlhb9	0.99	1.00	0.97	0.92	0.97	0.02	0.01	53.61	
Bmp15	-	-	-	-	-	-	-	0.22	
Car5b	0.48	0.77	0.28	0.45	0.59	0.44	0.35	1.18	
Cfp	0.89	NA	0.96	0.96	0.99	0.12	0.01	28.36	
Col4a5	-	-	-	-	-	-	-	1.81	
Cstf2	0.99	1.00	0.94	0.95	0.97	0.00	0.01	118.51	
Ctps2	0.99	1.00	0.96	0.95	0.98	0.01	0.01	113.44	
Cybb	0.30	0.54	0.93	0.84	0.93	0.00	0.01	291.67	
Ddx3x	0.79	0.99	0.73	0.91	0.72	0.02	0.02	299.47	
Dlg3	0.97	0.99	0.94	0.91	0.98	0.01	0.08	16.77	
Dynlt3	-	-	-	-	-	-	-	54.06	
Ebp	1.00	1.00	0.91	0.90	0.97	0.01	0.01	44.00	
Eif2s3x	**0.49**	**0.52**	**0.53**	**0.56**	**0.53**	**0.25**	**0.19**	**46.76**	**yes**
Ercc6l	1.00	1.00	0.90	0.89	0.98	0.01	0.10	21.35	
F8	0.88	1.00	-	0.76	1.00	-	0.33	3.55	
Fam199x	0.99	1.00	0.98	0.97	0.94	0.06	0.05	21.38	
Fam3a	0.99	1.00	0.94	0.96	0.97	0.01	0.01	27.94	
Fam50a	0.98	0.99	0.97	0.90	0.96	0.00	0.00	48.07	
Firre	0.93	0.93	-	-	-	0.00	-	8.42	

Flna	1.00	0.99	0.96	0.94	0.96	0.01	0.01	811.48	
Fmr1	0.99	1.00	0.93	0.94	0.98	0.10	0.03	93.35	
Ftx	0.91	0.91	0.96	0.92	0.95	0.35	0.21	25.17	
Fundc2	1.00	1.00	0.97	0.92	0.98	0.01	0.00	29.13	
G530011O06Rik	0.20	0.13	0.57	0.71	0.48	0.61	0.44	1.66	
G6pdx	0.99	1.00	0.96	0.95	0.97	0.01	0.01	73.63	
Gdi1	0.99	0.98	0.93	0.94	0.95	0.01	0.01	463.40	
Gla	0.97	1.00	-	0.84	1.00	-	-	5.44	
Gnl3l	0.99	0.99	0.95	0.92	0.94	0.03	0.03	64.96	
Gpm6b	0.58	-	0.93	0.79	0.95	0.23	0.24	3.03	
Gprasp1	0.98	0.99	0.93	0.90	0.94	0.01	0.01	51.27	
Gripap1	0.99	0.99	0.96	0.95	0.97	0.01	0.02	104.11	
Gk	0.95	0.99	0.97	0.89	0.99	0.01	0.10	15.92	
Hdac6	0.94	0.98	0.95	0.92	0.95	0.05	0.12	23.25	
Hs6st2	-	-	-	-	-	-	-	1.11	
Htatsf1	0.99	1.00	0.95	0.96	0.96	0.01	0.00	75.27	
Huwe1	0.98	0.99	0.92	0.93	0.95	0.02	0.03	100.25	
Idh3g	1.00	1.00	0.95	0.94	0.97	0.01	0.00	221.82	
Ids	1.00	0.99	0.97	0.92	0.96	0.01	0.02	47.79	
Ikbkg	0.99	1.00	0.96	0.95	0.97	0.01	0.01	85.87	
Il13ra1	-	-	-	-	-	-	-	1.32	
Iqsec2	-	-	-	0.58	-	-	-	1.14	
Irak1	0.99	0.99	0.95	0.94	0.97	0.01	0.01	144.37	
Itm2a	0.99	0.99	0.92	0.92	0.97	0.05	0.04	43.54	
Jpx	-	-	-	-	-	-	-	#DIV/0!	
Kdm5c	**0.60**	**0.63**	**0.62**	**0.59**	**0.61**	**0.29**	**0.28**	**175.94**	**yes**
Kdm6a	**0.55**	**0.55**	**0.66**	**0.58**	**0.58**	**0.40**	**0.39**	**145.05**	**yes**
Kif4	0.99	0.99	0.86	0.88	0.98	0.14	0.17	33.39	
Lamp2	0.98	1.00	0.93	0.95	0.94	0.01	0.02	105.10	
Maged1	0.96	1.00	0.98	0.95	0.96	-	0.00	4.83	
Magt1	0.99	0.99	0.94	0.94	0.97	0.01	0.01	70.50	
Mecp2	0.99	0.99	0.95	0.95	0.95	0.01	0.04	44.26	
Mid1	0.45	0.40	0.62	0.68	0.65	0.08	0.06	6.87	
Mid2	-	-	-	-	-	-	-	0.68	
Mmgt1	0.99	1.00	0.93	0.91	0.98	0.04	0.05	43.28	
Msn	0.99	1.00	0.94	0.94	0.97	0.01	0.01	1347.44	
Mtcp1	0.96	0.99	0.91	0.91	0.90	0.00	0.06	14.88	
Nkap	0.99	0.99	0.94	0.94	0.97	0.01	0.02	38.00	
Nudt11	-	-	-	-	-	-	-	0.08	
Ofd1	0.97	0.99	0.94	0.94	0.96	0.02	0.06	19.34	
Ogt	1.00	1.00	0.96	0.94	0.96	0.01	0.01	556.30	
Otud5	0.99	1.00	0.93	0.94	0.96	0.03	0.04	227.57	
Pbdc1	**0.65**	**0.77**	**0.75**	**0.71**	**0.80**	**0.30**	**0.23**	**53.17**	**yes**

Pcdh19	-	-	-	-	-	-	-	0.84	
Pdha1	1.00	1.00	0.93	0.96	0.96	0.01	0.02	107.61	
Pim2	0.99	0.99	0.83	0.88	0.95	0.01	0.01	55.97	
Plp1	1.00	0.97	0.89	0.92	0.99	0.23	0.12	2.59	
Pls3	-	-	-	1.00	-	-	-	2.18	
Pqbp1	0.99	0.99	0.95	0.90	0.96	0.01	0.01	104.39	
Rbbp7	-	-	-	-	-	-	-	192.46	
Rlim	0.98	1.00	0.93	0.96	0.96	0.01	0.01	84.62	
Rnf128	-	-	-	-	-	-	-	0.53	
Rps4x	0.99	1.00	0.95	0.94	0.96	0.01	0.01	196.71	
Sh3bgrl	0.99	0.99	0.93	0.94	0.96	0.01	0.04	134.77	
Shroom4	-	-	-	-	-	0.65	-	1.07	
Siah1b	0.89	0.98	0.95	0.90	0.97	0.02	0.00	7.53	
Slc16a2	-	-	-	-	-	-	-	0.50	
Slc35a2	0.99	1.00	0.96	0.96	0.97	0.02	0.02	36.33	
Slc6a8	-	-	-	-	-	-	-	0.66	
Snx12	0.99	0.99	0.98	0.95	0.97	0.23	0.19	42.85	
Ssxb3	-	-	-	-	-	-	-	0.15	
Suv39h1	0.99	1.00	0.88	0.93	0.97	0.05	0.14	62.48	
Syap1	0.99	1.00	0.99	0.96	0.97	0.01	0.00	35.80	
Syp	-	-	-	-	-	-	-	1.33	
Tab3	0.98	0.99	0.95	0.90	0.97	0.02	0.07	20.34	
Taf1	0.98	1.00	0.86	0.90	0.94	0.05	0.03	64.98	
5430427O19Rik	0.63	0.93	0.96	0.92	0.98	0.02	0.03	40.30	
Tmem164	0.99	1.00	0.88	0.90	0.96	0.01	0.02	88.06	
Tmem29	0.83	0.83	0.86	0.89	0.83	0.25	0.29	3.81	
Tmem47	-	-	-	0.20	-	-	-	0.74	
Tmsb15l	-	1.00	1.00	0.74	1.00	-	-	0.28	
Tmsb4x	1.00	1.00	0.96	0.96	0.97	0.00	0.00	2015.50	
Uba1	1.00	0.99	0.95	0.93	0.96	0.00	0.00	326.39	
Ubl4a	1.00	0.99	0.95	0.90	0.95	0.01	0.00	69.97	
Usp9x	0.94	0.99	0.95	0.92	0.95	0.05	0.04	80.36	
Utp14a	**0.99**	**0.62**	**0.94**	**0.77**	**0.77**	**0.03**	**0.22**	**26.22**	**yes**
Vbp1	1.00	1.00	0.95	0.96	0.98	0.01	0.03	121.16	
Vsig4	-	-	-	-	-	-	-	0.27	
Wdr13	0.99	1.00	0.95	0.95	0.96	0.03	0.00	43.80	
Xist	**0.01**	**0.01**	**0.10**	**0.09**	**0.07**	**0.96**	**0.99**	**162.43**	**yes**
Yipf6	0.99	0.99	0.96	0.97	0.97	0.05	0.08	43.46	
Zbtb33	1.00	-	-	-	-	-	-	5.64	
Zfp280c	0.96	0.99	0.90	0.92	0.94	0.03	0.06	26.07	
Zmym3	0.99	1.00	0.93	0.93	0.96	0.02	0.04	37.85	
Zrsr2	0.98	1.00	0.92	0.93	0.95	0.01	0.01	33.89	
Gm8822	**0.38**	**0.35**	**0.44**	**0.35**	**0.38**	**0.39**	**0.30**	**19.69**	**yes**