Allelic expression analysis of Imprinted and X-linked genes from bulk and single-cell transcriptomes

Paolo Martini[1§], Gabriele Sales[1§], Valentina Perrera[2,3] , Linda Diamante[2] , Chiara Romualdi[1*] and Graziano Martello[2*]

Affiliations:
[1] Department of Biology, University of Padova, Padua, Italy.
[2] Department of Molecular Medicine, Medical School, University of Padova, Padua, Italy.
[3] International School for Advanced Studies (SISSA/ISAS), Trieste, 34136, Italy.
*e-mail: chiara.romualdi@unipd.it; graziano.martello@unipd.it
§ Authors contribute equally to the work
†Current address: International School for Advanced Studies (SISSA/ISAS), Trieste, 34136, Italy.

Abstract

Genomic imprinting and X chromosome inactivation (XCI) are two prototypical epigenetic mechanisms whereby a set of genes is expressed monoallelically in order to fine tune their expression levels. Defects in genomic imprinting have been observed in several neurodevelopmental disorders, in a wide range of tumors and in induced pluripotent stem cells (iPSCs). Single Nucleotide Variations (SNVs) are readily detectable by RNA-sequencing allowing determination of whether imprinted or X-linked genes are aberrantly expressed from both alleles, although standardised analysis methods are still missing. We have developed a tool, named BrewerIX, that provides comprehensive information about allelic expression of a large, manually-curated set of imprinted and X-linked genes. BrewerIX does not require programming skills, runs on a standard personal computer, and can analyse both bulk and single-cell transcriptomes of human and mouse cells directly from raw sequencing data. BrewerIX confirmed and extended previous observations regarding the aberrant expression of imprinted genes in pluripotent cells, in the early embryo and in breast cancer cells and identified new genes escaping XCI in human somatic cells. We believe BrewerIX will be useful for the study of genomic imprinting and XCI during development and reprogramming, and for detecting aberrations in cancer and iPSCs. Due to its ease of use to non-computational biologists, its implementation could become standard practice during sample assessment, thus raising robustness and reproducibility of future studies.

Main

Gene imprinting is used to control the dosage of a specific set of genes (imprinted genes) by selectively silencing one of the two copies of the gene (either the maternal or the paternal allele). In female cells, also the genes on the X chromosome are expressed monoallelically thanks to a random epigenetic silencing mechanism called X chromosome inactivation (XCI). X-linked and imprinting diseases are the most common congenital human disorders, because loss-of-function mutations in the single expressed allele will not be buffered by the second silenced allele[1]. Imprinted genes were initially isolated as regulators of fetal growth and their aberrant expression has been related to cancer[2–4]. For these reasons, analysing the imprinting and XCI status is crucial in many fields including cancer research, regenerative medicine and assisted reproductive technology.

Correct imprinting information and reactivation of X chromosome are criteria used to evaluate the quality of induced pluripotent stem cells (iPSCs). Although iPSCs hold the promise for effective approaches in regenerative medicine, disease modelling and drug screening (for review see Perrera and Martello[5]), their safety is compromised by frequent genetic and epigenetic aberrations, such as Loss of Imprinting (LOI) or a variable X chromosome status[6–15].

Allelic expression can be determined by the presence of Single Nucleotide Variants (SNV) in RNA-sequencing (RNAseq) data. However, at the time of writing, no standardised pipelines for analysis of allelic expression of Imprinted and X-linked genes have been developed. Existing pipelines use different combinations of tools and rely on different parameters that were set to analyze specific data and to address specific questions[15–17]. Moreover, these pipelines need skilled bioinformaticians to be run. A complete and easy to use tool, which does not require programming skills, is still missing.

Motivated by this need, we built BrewerIX, an app available for macOS and Linux Systems that looks for bi-allelic expression of experimentally validated imprinted genes (see Supplementary Table 1 and 2 for a manually curated list of human and mouse genes) and genes on the sex chromosomes. Bi-allelic expression of imprinted genes will indicate LOI. Bi-allelic expression of X-linked genes may indicate reactivation of the X chromosome, as expected in the early embryo[18] or in naive pluripotent stem cells[12,13,19,20], X chromosome erosion, as

observed after extensive culture of pluripotent cells[21], or simply escape of single genes from the XCI mechanisms, as recently documented in somatic cells[22,23].

BrewerIX (freely available at https://brewerix.bio.unipd.it) takes as input either bulk or single-cell RNAseq data, analyzes reads mapped over the SNV distributed on imprinted genes, X chromosome and Y chromosome and generates imprinting and XCI profiles of each sample displaying them in an intuitive way.

BrewerIX implements three pipelines with different aims (Fig. 1a, Supplementary Fig. 1). The Standard pipeline is meant to rapidly have the imprinting and X inactivation status of a set of samples (Fig. 1a). Here, BrewerIX will align each sample, filter alignments and call Allele Specific Expression (ASE) Read counter (see Methods) using a set of pre-compiled bi-allelic SNVs. Before visualization, SNVs are collapsed by genes to create a table that is displayed by the user interface. The Complete pipeline sacrifices speed for the sake of completeness by using a larger set of SNVs, while the Tailored pipeline uses a specific set of SNVs that the user might detect from DNA-seq data (Supplementary Fig. 1).

The end-point of the pipelines is a table that is visualized by the user interface (UI). The UI presents the results using two graphical panels. The gene summary panel shows a matrix of dots with as many rows as the number of genes and as many columns as the number of samples analyzed. The size and the color of the dot is proportional to the confidence of our estimate: i) the larger the dot, the higher the number of SNVs supporting our estimate; ii) the brighter the color, the closer to 1 is the average of the allelic ratios (minor/major) of all bi-allelic SNVs. Empty dots are genes with no evidence of bi-allelic expression. Lack of a dot means that no reads overlapped the SNVs available for the gene, indicating that the gene was not detected.

The SNVs summary panel shows a set of barplots (one set for each sample) with as many bars as the number of SNVs per gene. Here blue is the color of the reference allele and red is the alternative/minor one. Solid colors indicate bi-allelic SNVs, transparent colors indicate mono-allelic SNVs, while those SNVs that do not meet the minimal coverage are shown in gray. When a gene shows no evidence of any genuine bi-allelic SNVs, we collapse the counts over a virtual SNV (named “rs_multi”) to give an indication of its expression.

The UI allows to set different filters according to the dataset features, based on the following 4 parameters:

1. the overall depth, representing the number of reads mapping on a given SNV;
2. the minor allele count, indicating the absolute number of reads mapping on the less frequent SNV variant among the two detected (i.e. the minor allele);
3. the threshold to call a bi-allelic SNV, which can be either a cutoff on the allelic ratio (AR, minor/major allele) or the p-value of a binomial test;
4. the minimal number of bi-allelic SNVs needed to call a bi-allelic gene, based on the assumption that when a gene is expressed bi-allelically, multiple bi-allelic SNVs should be detected.

Default values of the parameters have been empirically selected to minimize the number of false positives. A false positive call is a SNVs not present in the DNA, detected only at the RNA level due to sequencing and caller errors. We reasoned that false positive calls can be estimated using genes on sex chromosomes, of whom only a single allele is present. Thus, we analyzed bulk RNAseq samples of 6 normal male BJ fibroblasts from 3 published datasets (see Supplementary Table 3, describing all datasets used in this study). We collected on sex chromosomes all the SNV with an overall depth of >=5 reads in at least one sample. Then we evaluated the false positive ratio at increasing thresholds (minor allele count and overall depth) as the ratio between the number of bi-allelic SNVs (those with an AR >= 0.2 as in [15] and [24]) over the total number of SNVs detected in the sample (those with more than 5 reads).
As shown in Fig. 1b, the frequency distribution of false positive rate shows a clear elbow point at a minimal coverage of 20 reads and 4 reads for the minor allele finding an average of 2 false positive calls every 10^5 SNVs analyzed. No biallelic SNVs were detected in Y in any of the analyzed samples. Looking at the distribution of the false positive SNVs in the X chromosome, we found no correlation between the calls and the number of tested SNV (Supplementary Fig. 2).

To gain further confidence in methods based on RNAseq data, we calculated the number of false positive calls detected by SNP-array, a technique specifically developed and extensively used to detect SNVs. We analyzed genomic DNA from BJ fibroblasts profiled with Affymetrix Mapping 250K Nsp SNP Array, and we found that the number of false positives detected was

100 times higher (2 every 10^3 evaluated SNVs, Supplementary Fig. 3) confirming that RNAseq data is more accurate in detecting allelic imbalance.

Although the defined thresholds minimize false positives, we investigated their power of detecting actual bi-allelic genes. For this reason, we analyzed the pseudo-autosomal region 1 (PAR1), a short region of homology between the X and Y chromosomes which behaves like autosomes and contains 22 genes expressed bi-allelically. Our results indicate that the number of bi-allelic calls in PAR1 is significantly higher (p=0.031 Wilcoxon signed rank test) than the mean number of false positives detected in the remaining part of the X chromosome (Supplementary Fig. 2).

To further test the capacity to detect actual bi-allelic expression, we analyzed RNAs-seq data from female human naive iPSCs (HPD08 - GSM2988908), bearing two active X chromosomes[13,20]. We detected 104 bi-allelic genes on the entire X chromosome out of 382 detected genes. Overall we conclude that the chosen parameters allow detection of bi-allelic expression while minimising false-positive calls.

Our default parameters for standard bulk RNAseq samples (>10M reads/sample) are 20, 4 and 0.2 for overall depth, minor allele count and AR respectively. Additionally, we call a gene bi-allelic when at least 2 bi-allelic SNVs are detected, in order to filter out potential sequencing artifacts. To test BrewerIX functionalities we analyzed 8 datasets, including both bulk and single cell RNAseq, different organisms (human and mouse) and different biological systems (iPSCs, cancer cells, early embryonic development).

Reprogramming of human somatic cells to pluripotency is associated with imprinting abnormalities[5], both in the case of conventional, or "primed", iPSCs and in the case of naive iPSCs[6–8,13,15,25–27]. We analyzed 10 isogenic bulk RNAseq samples, including 6 BJ fibroblast, 1 primed iPSC line and 3 naive iPSCs. We run the analysis both in Complete mode (Fig. 1c) and Standard mode (Supplementary Fig. 4), obtaining highly comparable results. MEG3 showed bi-allelic expression specifically in naive iPSCs (Fig. 1c-d), as previously reported[13,28]. Several other imprinted transcripts, such as H19, MEG8, INPP5F and NLRP2 showed bi-allelic expression in naive iPSCs.

To experimentally validate these results and further demonstrate the accuracy of the default parameters, we performed Sanger sequencing after PCR amplification of genomic DNA from

1 naive iPSC line and confirmed the presence of 12 randomly selected SNVs (Supplementary Table 4 and Fig. 1e), while bi-allelic expression of MEG3 was confirmed in 3 independent naive iPSC lines (Fig. 1e). A second dataset of human fibroblasts (HFF) and matching naïve iPSCs (HPD06[13]) was analyzed in Standard mode, confirming bi-allelic expression of H19, MEG3, INPP5F and NLRP2 only in naive cells (Supplementary Fig. 5), as previously reported[13,19,28].

We analyze a dataset of murine Embryonic Stem cells (mESCs) expanded under different culture conditions. Yagi and colleagues reported that expanding mESCs in 2i/L conditions resulted in LOI, while mESCs in S/L conditions mostly retained correct imprinting[29]. With BrewerIX we obtained identical results for 7 out of 8 imprinted genes analyzed by Yagi and colleagues (Fig. 1f). The eighth gene, Zim2, showed a too low overall sequencing depth and was not analyzed. We conclude that BrewerIX detected LOI events in both human and mouse naive pluripotent stem cells from bulk RNAseq data, in agreement with previous analyses[13,28,29].

Next, we wanted to compare the performance of BrewerIX on matching bulk and single-cell RNAseq data. Using bulk samples from mESCs cultured in 2i/L or S/L conditions[30], we identified 19 LOI events, with Ddc showing LOI specifically in 2i/L and Gatm, Pon2 and Blcap showing LOI only in S/L (Fig. 1g).

We then analyzed single-cell data (384 cells from 2i/L and 288 from S/L) using our default parameters, considering a gene bi-allelically expressed when a single SNVs was found bi-allelic in at least 20% of cell analyzed expressing such gene (Fig 2a). We observed that those genes with multiple bi-allelic SNVs in bulk analysis, such as Impact, Lin28a, and Inpp5, were found bi-allelic also in a large fraction (>50%) of single cells analyzed. Several LOI events were detected only in bulk samples, possibly because single-cell RNAseq detects preferentially the 3' end of transcripts, limiting the number of SNVs detected. Despite such limitation, some bi-allelic genes could be detected only by single-cell RNAseq (e.g. Ccdc40, Peg10 and Plagl1), indicating that only single-cell RNAseq allows the detection of LOI events occuring in a limited fraction of cells.

Deng and colleagues analyzed the gene expression of single cells from oocyte to blastocyst stages of mouse preimplantation development describing that in female embryos the paternal

X chromosome is activated beyond the four-cell stage and subsequently silenced[30]. BrewerIX results were highly concordant with those generated with a custom pipeline by Deng and colleagues, confirming the transient reactivation of the paternal X chromosome (Fig. 2b and Supplementary Fig. 6). Next, we observed an expected monoallelic expression of most of the imprinted genes (Fig. 2c and Supplementary Fig. 7), although few of them, such as Gnas, Nap1l4, Cd81, Usp29, showed bi-allelic expression at several stages of pre-implantation embryos, suggesting that imprinted expression might be consolidated later in development.

Next, we analyzed a human somatic single-cell RNAseq dataset[17] and observed that 38 genes showed bi-allelic expression in 20% of cells (Fig. 2d). Only 4 of these genes (GLIS3, GNAS, ATP10A and TFPI2) were also found bi-allelic by the authors of the original study[17]. We extended the analysis to X-linked genes and found that, out of 608 detected genes, 35 genes escaped XCI in at least two individuals (Fig. 2e). Notably, only 16 out of 35 (45%) were previously identified as escapees[31]. We conclude that BrewerIX efficiently identifies LOI and XCI escape events occurring in small fractions of somatic cells from single-cell transcriptomes.

Different cancers, such as breast, kidney and lung, are characterized by frequent expression level changes of imprinted genes, often accompanied by DNA methylation level changes in several imprinted domains, such as the PEG3, MEST and GNAS[32]. To test whether BrewerIX could detect LOI events in cancer cells, we analyzed 515 single cell samples and matching bulk samples from 11 breast cancer patients[33].
We first defined what SNVs could be detected in bulk samples from patients using BrewerIX in Complete mode. Such SNVs list was then used to interrogate in Tailored mode the single-cell dataset, in which the authors classified the cells as Tumor and non-Tumor (i.e. stromal and immune cells surrounding the tumor).

From bulk RNAseq data, we found that 7 imprinted genes showed bi-allelic expression, among them DNMT1 and GNAS were detected in at least 4 patients (Fig. 2f).
DNMT1 and GNAS were also found bi-allelic in the single-cell dataset, in a high percentage of both tumor and non-tumor cells (Fig. 2g). We detected 18 additional bi-allelic genes, including MEST and OSBPL5 that showed bi-allelic expression specifically in tumor cells. Such results indicate that single-cell analyses outperform bulk analyses in the case of

heterogeneous cancer samples and that imprinting abnormalities might be much more widespread in cancer cells than currently thought.

The results obtained by BrewerIX on the selected case studies outcompeted published custom pipelines confirming and extending published results, demonstrating the reliability and usefulness of the tool. For the analysis of relatively homogeneous cell populations, such as pluripotent cells in culture, we conclude that bulk RNAseq data allowed robust identification of LOI events. Conversersely, when heterogeneous populations of cells, such as cancer samples, are analyzed, only single-cell measurements allowed to detect widespread events of LOI or XCI escape, indicating that such phenomena might have been underestimated for technical limitations.

Due to the ease of use of BrewerIX to non-computational biologists, we believe that its implementation could become standard practice during assessment of newly generated pluripotent cells, as well as for the study of the molecular mechanisms underlying genomic imprinting and XCI, hopefully raising robustness and reproducibility of future studies.

Methods

General overview

BrewerIX is implemented as a native graphical application for Linux and macOS. Upon installation, BrewerIX automatically downloads all the required software dependencies and data for both the human and mouse species, which are then cached for later usage. An intuitive user interface guides the user to configure the analysis.

BrewerIX requires a directory where the FASTQ files are stored. All the FASTQs will be processed and all of them need to have homogeneous read layouts (all single-end or all paired-end). After choosing the appropriate read layout, the user can choose among three analysis modes: Standard, Complete and Tailored.

The Standard and Complete mode run with pre-compiled set of SNVs: bi-allelic and bi-allelic plus multi-allelic respectively (see "Precompiled sets of SNVs" paragraph for details). The Tailored mode requires a user-defined set of SNVs. Finally, the user has the option of exploiting multiple processing cores on his system to speed-up the analysis. All the three analysis modes end up saving a "brewer-table" and opening the user interface to browse results.

The results are organized by the user interface in tabs named "Imprinted Genes", "Chromosome X" and "Chromosome Y" that give access to the respective analyses. All the tabs are organized similarly: the left panel provides a gene summary, while the right panel displays per-gene SNV details.

The gene summary has samples in the columns and genes on the rows. Position of the samples can be arranged just dragging them in the correct order. Genes on the rows are sorted according to their genomic position (chromosome and transcription start site). The circles represent the summarized allelic ratio for each gene in the sample: the size of the circle reflects the number of SNVs used to compute the average allelic ratio.

Default settings can be easily changed using the "Options" menu. "Filter SNVs" controls tunable parameters related to SNVs: overall depth, minimal number of minor allele count, number of bi-allelic SNVs to call a gene as bi-allelic . Moreover, the user can choose to call a bi-allelic SNV based on a fixed threshold or a binomial test. "Filter genes" allows the user to choose the set of genes to display: all, only those detected (i.e. those with a sufficient overall depth) or only those genes that are bi-allelic in at least one sample that was analyzed. Additionally, the user can control the source of imprinted genes to be included in the analysis: human and mouse have 3 sources that can be combined (see "Knowledge base" paragraph).

Finally, the user can control the allelic ratio measure, as either minor allele / major allele, or minor allele / total counts. Clicking on the gene names or on the dots will open (or update) the right-hand panel that will display the SNVs used to perform the call. Each SNV is represented with a bar where reference allele counts are in blue and alternatives in red (see SNV calling section). An empty bar means no reads are available for that SNV. Gray bars indicate SNVs that do not reach the overall depth to be considered detected. Bright colors indicate that the SNV is bi-allelic; dim colors indicate a SNV that is not bi-allelic. Both left and right panels can be saved as PDF files. Moreover, the right panel (i.e. the gene summary) can be exported as a tab-delimited file to allow further analysis. All exports reflect the filters chosen.

Implementation Details:

User Interface

The BrewerIX graphical interface is distributed as a native application for both Linux and macOS. It is written in the Haskell programming language and makes use of the wxWidgets cross-platform GUI library. Plots are generated using the Cairo library and its PDF output capabilities. The Linux version of the application is packaged using the AppImage tool.

Core Computational Pipeline

The computational pipeline is implemented in Python and is available as a Python package called brewerix-cli at github.com/Romualdi-Lab/brewerix-cli. The pipeline performs the alignment, allelic count and creation of the result table called “brewer-table”. The pipeline can be run also using the command line interface (CLI) implemented by brewerix-cli itself. The final output of the CLI is the “brewer-table” that is parsed by the visual interface to produce the BrewerIX visual outputs. The CLI has been though for advanced users willing to analyze their own set of genes or genomes of different species. The minimum required inputs are the following: a genome (fasta format) and its index for hisat2, genome dict (computed with GATK) and genome fasta index, a bed file indicating the region of interest (i.e. imprinted genes and genes on the sex chromosomes), a set of bi-allelic SNVs with reference alleles that must be present in the reference genome.

Knowledge base

We manually curated a comprehensive set of imprinted genes from different sources. For human and mouse imprinted genes, we collected the data from the Geneimprint database

(http://geneimprint.com/) and Otago database (http://igc.otago.ac.nz/home.html). We excluded all genes labeled as "predicted" or "notImprinted" and manually curated "conflicting data". We added human imprinted genes identified by Santoni and colleagues[17] (https://doi.org/10.1016/j.ajhg.2017.01.028) and mouse imprinted genes regulated by H3K27me3 in the early embryo, identified by Inoue and colleagues[34] (https://doi.org/10.1038/nature23262). The "Gene filters" command on the User interface allows choosing any combination of these resources. The manually curated gene lists are shown in Supplementary Tables 1 and 2.

The manually curated list of imprinted genes, together with genes on X and Y chromosomes, are the starting point to build the Knowledge base. Upon first usage, BrewerIX downloads the pre-built species-specific Knowledge base. This task needs to be done only once.

The Knowledge base contains the genome, the genome index directory, the bi-allelic SNV file, the multi-allelic SNV file, the regions with the genes of interest.

To create a custom Knowledge base, we implemented a Python package called brewerix-prepare-knowledgebase that is able to create a knowledge base for brewerix-cli from the ENSEMBL database. Minimal inputs are the species (must be a valid ensembl species), the chromosomes and a list of the genes of interest (tested on ENSEMBL 98 for mouse and human).

Precompiled sets of SNV

BrewerIX comes with two precompiled sets of SNVs: bi-allelic set and a multi-allelic set of SNVs for both human and mouse. SNVs were downloaded from ENSEMBL variants (annotation version 98). We removed INDELs and the SNVs whose reference alleles differed from the reference genome. Bi-allelic SNVs and multi-allelic SNVs were assigned to the bi-allelic and multi-allelic set accordingly.

Alignments

BrewerIX requires fastq files as input. The pipeline works with homogeneous library layout i.e. all fastqs are either single or paired end. The fastq files are aligned to a reference genome. The user can choose between Mouse GRCm38.p6 or human GRCh38.p13 genome. Alignments are performed using Hisat2 (version 2.1.0, default parameters) and filtered to keep only reads laying on genes of interest.

SNV calling

SNVs are called only at multi allelic SNVs using HaplotypeCaller from GATK v4.1. Calls are performed as if all the samples have the same genotype, i.e. all in the same batch. The reference and the most represented alternative allele are selected. We set the following parameters: "--max-alternate-alleles 1 -stand-call-conf 1 --alleles multi_allele_vcf_file --dbsnp multi_allele_vcf_file".

Allelic count

Allelic count is performed using ASEReadCounter with default parameters from GATK v4.1. This tool, given a set of loci and a bam file, allows computing the reads bearing the reference and the alternative allele. Sample-specific results are collapsed into an ASER table.

The brewer-table

The brewer-table is created by the core computational pipeline and contains all the SNVs that were detected by at least 5 reads. To reduce the table size, genes without any alternative allele in any of the detected SNVs are collapsed into a meta-SNV that gives an indication about the coverage of the gene under analysis.

False-positive detection evaluation

To evaluate BreweIX performance we first estimated the number of false positives. We analyze 6 male BJ fibroblasts looking for how many SNVs are called bi-allelic in the X chromosome. We reasoned that all the bi-allelic SNVs called outside the Pseudoautosomal Regions (PARs) are false positives, given that male cells have only one copy of the X chromosome.

We considered all SNVs on the sex chromosomes with an overall depth of >=5 reads in at least one sample. A SNV is biallelic when its allelic ratio >= 0.2 (minor/major) and we increased the allele count and overall depth to find the optimal cut. We found a clear elbow point at a minimal coverage of 20 and 4 reads for the minor allele finding an average of 2 false positive calls every 10^5 SNVs analyzed.

Although such value appeared low, we wanted to compare it against the false-positive rate obtained with an independent technique that has been developed and used for analysis of SNVs, SNP-arrays.

We compared the number of SNVs in X chromosome from RNAseq to the number of SNVs in the X chromosome detected by Affymetrix SNP-array using BJ DNA sample (GEO accession GSE72531) and obtained 2 false-positive bi-allelic SNVs every 10^3 evaluated SNVs.

Case Studies

We chose as case studies 7 datasets that were very diverse, in order to fully exploit all the features of the tools we have developed. In the following sections we will describe in detail the workflow of each dataset. All RNAseq data but one were downloaded from GEO database using fastq-dump from sra-tools version 2.8.2. Only mouse ESCs dataset was downloaded from Array Express via direct link.

All datasets images were created using the BrewerIX-core imprinted genes (i.e. genes curated from geneimprint DB and Otago) unless stated otherwise. Moreover images were created showing only "significant" genes with default parameters i.e minimal coverage of 20, 4 reads for the minor allele, allelic ratio (mino/major) >= 0.2 and at least two or one biallelic SNV per gene in bulk and single cell sequencing respectively.

BJ fibroblast dataset.

We collected BJ RNAseq data from 3 sources on the GEO database: GSE110377 (BJ fibroblast GSM2988896; primed iPSC GSM2988902, naive iPSC GSM2988898, GSM2988903, GSM2988904), GSE126397 (BJ fibroblasts GSM3597749 and GSM3597750) and GSE63577 (BJ fibroblasts GSM1553088-GSM1553090). To deal with the heterogeneous read layout (single and paired-end) of the sequencing data, we aligned each batch to the reference human genome using hisat2, with default parameters. We use BrewerIX-cli to run the analysis starting from the alignment files (bams). We used the "complete" analysis mode and loaded the "brewer-table" on the visual interface to explore the results.

HFF dataset

HFF normal samples were downloaded from GSE93226 (GSM2448850-GSM2448852) while reprogrammed iPSC from GSE110377 (GSM2988900). As for the BJ fibroblast dataset, we computed single and paired end alignments separately (hisat2, default parameter) and then run brewerix-cli in "standard" mode. Panels summarizing the results have been generated with BrewerIX's User interface.

Yagi et al. dataset - mouse ESCs.

Yagi dataset (GEO accession GSE84164; GSM2425488-GSM2425495) was fully analyzed by brewerIX with the Complete analysis mode. To generate the figure, we selected from the brewer-table only those genes shown in Yagi et al. and submitted the new table back to BrewerIX.

Kolodziejczyk et al. / Kim et al. dataset - mouse ESCs

In this dataset, we analyzed mES cells cultured in 2i/L or S/L downloaded from Array Express under the accession E-MTAB-2600. We analyzed three bulk samples (one cultured in 2i/L and two in S/L) and 682 single cell samples (384 cultured in 2i/L and 288 in S/L).

Both bulk and the single cell RNAseq datasets were analyzed using BrewerIX in Standard mode. Bulk data visualization on the three samples was performed using BrewerIX User interface.

Single cell RNAseq results were visualised using custom R code available at github.com/Romualdi-Lab/. Results were summarized by the two categories: 2i/L and S/L. We analyzed genes that are expressed in at least 10 cells in at least one category. We considered a gene bi-allelically expressed when at least one SNV was found bi-allelic in at least 20% of cells analyzed expressing such gene (other parameters remain default).

Deng et al. dataset - oocyte to blastocyst

Single cell RNAseq dataset were downloaded from GEO accession GSE45719 (GSM1112490-GSM1112581 and GSM1112603-GSM1278045; female samples include GSM1112504-GSM1112514, GSM1112528-GSM1112539, GSM1112543-GSM1112553, GSM1112626-GSM1112640, GSM1112656-GSM1112661, GSM1112696-GSM1112697, GSM1112702-GSM1112705; male samples include GSM1112490-GSM1112503, GSM1112515-GSM1112527, GSM1112540-GSM1112542, GSM1112554-GSM1112581, GSM1112611-GSM1112625, GSM1112641-GSM1112653, GSM1112654-GSM1112655, GSM1112662-GSM1112695, GSM1112698-GSM1112701, GSM1112706-GSM1112765; for remaining samples no sex specification were available). Analysis has been carried out using BrewerIX in "standard" mode. The computed values were used for downstream custom analysis (code can be found at github.com/Romualdi-Lab/).

For the X chromosome, we performed the analysis plotting the average of the allelic ratios in each developmental stage for male and female samples. We used developmental stages where

both male and female samples were present. Thus, we considered 4 male, 6 female in middle 2-cell (mid2cell); 4 male, 6 female for late 2-cell (late2cell); 3 male, 11 female for 4-cell (4cell); 27 male, 23 female for 16-cell (16cell); 28 male, 15 female for early blastocyst (earlyblast). To evaluate the performance of BrewerIX in detecting paternal-X chromosome re-activation, we downloaded Deng's processed dataset from the supplementary material of the manuscript[30]. To avoid any bias, we analyzed genes shared by Deng's processed dataset and BrewerIX generated data.

For imprinted genes, we plotted the Average Allelic Ratio (AAR) for each gene in each developmental stage. We analyzed the following developmental stages: 4 zygotes (zy), 8 early 2cell (early2cell), 12 middle 2 cell (mid2cell), 10 late 2cell (late2cell), 14 4-cell (4cell), 47 8-cell (8cell), 58 16-cell (16cell), 43 early blastocyst (earlyblast), 60 middle blastocyst (midblast), 30 late blastocyst (lateblast) and 10 fibroblast.

Santoni et al. / Garieri et al. dataset - human somatic cells

We used available data from 772 human fibroblasts (we analyzed 229, 159, 192 and 192 for IND1, IND2, IND3 and IND4 respectively) and 48 lymphoblastoid (IND5) cells from 5 female individuals (GEO accession GSE123028, GSM3493332-GSM3494151).

Single-cell RNAseq dataset was analyzed using BrewerIX in standard mode. The single cell RNAseq visual reports were produced with custom R code available at github.com/Romualdi-Lab/.

Results were summarized by individuals. We analyzed genes that are expressed in at least 10 cells in at least four categories. We considered a gene bi-allelically expressed when at least one SNV was found bi-allelic in at least 20% of analyzed cells that express that gene (other parameters remain default).

Chung dataset - Breast cancer

Chung and colleagues[33] analyzed 11 patients representing four subtypes of breast cancer (luminal A - BC01 and BC02, luminal B - BC03, HER2+ - BC04, BC05 and BC06 or triple negative breast cancer - TNBC – BC07-11). They obtained 515 single cell transcriptome profiles and 12 matched samples with bulk RNAseq from 11 patients (GEO accession GSE75688 all the samples listed in GSE75688_final_sample_information.txt.gz; B03 has both primary breast cancer and lymph node metastases). Bulk samples from the breast cancer dataset were analyzed using BrewerIX in Complete mode. Visual inspection was performed using

BrewerIX. The single-cell RNAseq dataset was run using the tailored mode with the SNV file created from analysis with Complete mode of the bulk matching RNAseq data. The single-cell RNAseq visual reports were produced with custom R code.

Cell sample annotations were downloaded from GEO database. Results were summarized by patients and according to available annotation further divided into tumor and non-tumor class (i.e. stromal and immune cells surrounding the tumor). Patients were included in the analysis if profiled for at least 8 tumor and 8 non-tumor cells. The numbers of cell analyzed for each patient class combination are the following: BC03_nonTumor=18, BC03_Tumor=15, BC03LN_nonTumor=43, BC03LN_Tumor=10, BC04_nonTumor=8, BC04_Tumor=47, BC06_nonTumor=10, BC06_Tumor=8, BC07_nonTumor=24, BC07_Tumor=26, BC07LN_nonTumor=26, BC07LN_Tumor=26.

We analyzed genes that were expressed in at least 2 cells in at least six categories. We considered a gene bi-allelically expressed when at least one SNV was found bi-allelic in at least 20% of analyzed cells that express that gene (other parameters remain default). Code to reproduce the figure can be found at github.com/Romualdi-Lab/ as well.

SNP detection via PCR followed by Sanger sequencing

Genomic DNA (gDNA) was extracted from cellular pellet with Puregene Core Kit A (Qiagen) according to the manufacturer's protocol; 1µg gDNA was used as a template for PCR using the Phusion High-Fidelity DNA polymerase (NEB, cat. M0530L).

Total RNA was isolated from cellular pellet using a Total RNA Purification kit (Norgen Biotek, cat. 37500), and complementary DNA (cDNA) was generated using M-MLV Reverse Transcriptase (Invitrogen, cat. 28025-013) and dN6 primers (Invitrogen) from 1000 ng of total RNA following the protocols provided by the manufacturers, including a step of TurboDNAse treatment (Thermo Scientific). cDNA was diluted 1:5 in water and used as a template for PCR using the Phusion High-Fidelity DNA polymerase; gDNA and cDNA were amplified by PCR using primers detailed in the Supplementary Table 5. PCR was conducted with the following program: denaturation at 98°C for 30s; 35 cycles of denaturation at 98°C for 10 s, annealing at temperature depending on primer sequence (T_m–5°C) for 30 s, elongation at 72°C for 15 s; final elongation at 72°C for 10 min.

PCR reaction products were resolved and imaged by agarose gel electrophoresis. The remaining PCR products were purified using the QIAquickPCR purification kit (Qiagen, cat. 28106) and direct sequencing was performed using the same primers used for PCR

amplification. Each PCR region of interest was sequenced at least twice, using both forward and reverse primers. Sanger sequencing was performed by Eurofins Genomics (https://www.eurofinsgenomics.eu/en/custom-dna-sequencing/gatc-services/lightrun-tube/). Sequence analysis and peak detection were performed using freely available ApE software (https://jorgensen.biology.utah.edu/wayned/ape/).

Data Availability

All RNAseq data used in this study were publicly available and obtained from either the Gene Expression Omnibus (GEO) database under the accession codes GSE110377; GSE126397; GSE63577, GSE110377; GSE126397; GSE63577 GSE93226; GSE110377 GSE84164 GSE123028 GSE45719 GSE75688 or from Array Express under the accession code E-MTAB-2600. Additional details about all datasets used in the study are in Supplementary Table 3. The raw Sanger sequencing data file underlying Fig. 1e and Supplementary Table 4 are provided as a Source Data file.

Code availability

BrewerIX is freely available for academic users at https://brewerix.bio.unipd.it and all code and tutorials are available at https://github.com/Romualdi-Lab/brewerix-cli under AGPL3 licence.

References

1. Monk, D., Mackay, D. J. G., Eggermann, T., Maher, E. R. & Riccio, A. Genomic imprinting disorders: lessons on how genome, epigenome and environment interact. *Nat. Rev. Genet.* **20**, 235–248 (2019).
2. Peters, J. The role of genomic imprinting in biology and disease: an expanding view. *Nat. Rev. Genet.* **15**, 517–530 (2014).
3. Kalish, J. M., Jiang, C. & Bartolomei, M. S. Epigenetics and imprinting in human disease. *Int. J. Dev. Biol.* **58**, 291–298 (2014).
4. Goovaerts, T. *et al.* A comprehensive overview of genomic imprinting in breast and its deregulation in cancer. *Nat. Commun.* **9**, 1–14 (2018).
5. Perrera, V. & Martello, G. How Does Reprogramming to Pluripotency Affect Genomic Imprinting? *Front. Cell Dev. Biol.* **7**, (2019).
6. Hiura, H. *et al.* Stability of genomic imprinting in human induced pluripotent stem cells. *BMC Genet.* **14**, 32 (2013).
7. Johannesson, B. *et al.* Comparable Frequencies of Coding Mutations and Loss of Imprinting in Human Pluripotent Cells Derived by Nuclear Transfer and Defined Factors. *Cell Stem Cell* **15**, 634–642 (2014).
8. Ma, H. *et al.* Abnormalities in human pluripotent cells due to reprogramming mechanisms. *Nature* **511**, 177–183 (2014).
9. Tchieu, J. *et al.* Female Human iPSCs Retain an Inactive X Chromosome. *Cell Stem Cell* **7**, 329–342 (2010).
10. Anguera, M. C. *et al.* Molecular Signatures of Human Induced Pluripotent Stem Cells Highlight Sex Differences and Cancer Genes. *Cell Stem Cell* **11**, 75–90 (2012).
11. Kim, K.-Y. *et al.* X Chromosome of Female Cells Shows Dynamic Changes in Status during Human Somatic Cell Reprogramming. *Stem Cell Rep.* **2**, 896–909 (2014).
12. Cantone, I. & Fisher, A. G. Human X chromosome inactivation and reactivation: implications for cell reprogramming and disease. *Philos. Trans. R. Soc. B Biol. Sci.* **372**, 20160358 (2017).
13. Giulitti, S. *et al.* Direct generation of human naive induced pluripotent stem cells from somatic cells in microfluidics. *Nat. Cell Biol.* **21**, 275–286 (2019).
14. Rugg-Gunn, P. J., Ferguson-Smith, A. C. & Pedersen, R. A. Epigenetic status of human embryonic stem cells. *Nat. Genet.* **37**, 585–587 (2005).

15. Bar, S., Schachter, M., Eldar-Geva, T. & Benvenisty, N. Large-Scale Analysis of Loss of Imprinting in Human Pluripotent Stem Cells. *Cell Rep.* **19**, 957–968 (2017).
16. Reinius, B. *et al.* Analysis of allelic expression patterns in clonal somatic cells by single-cell RNA–seq. *Nat. Genet.* **48**, 1430–1435 (2016).
17. Santoni, F. A. *et al.* Detection of Imprinted Genes by Single-Cell Allele-Specific Gene Expression. *Am. J. Hum. Genet.* **100**, 444–453 (2017).
18. Moreira de Mello, J. C., Fernandes, G. R., Vibranovski, M. D. & Pereira, L. V. Early X chromosome inactivation during human preimplantation development revealed by single-cell RNA-sequencing. *Sci. Rep.* **7**, 1–12 (2017).
19. Tomoda, K. *et al.* Derivation conditions impact x-inactivation status in female human induced pluripotent stem cells. *Cell Stem Cell* **11**, 91–99 (2012).
20. Sahakyan, A. *et al.* Human Naive Pluripotent Stem Cells Model X Chromosome Dampening and X Inactivation. *Cell Stem Cell* **20**, 87–101 (2017).
21. Xie, P. *et al.* The dynamic changes of X chromosome inactivation during early culture of human embryonic stem cells. *Stem Cell Res.* **17**, 84–92 (2016).
22. Cotton, A. M. *et al.* Analysis of expressed SNPs identifies variable extents of expression from the human inactive X chromosome. *Genome Biol.* **14**, R122 (2013).
23. Tukiainen, T. *et al.* Landscape of X chromosome inactivation across human tissues. *Nature* **550**, 244–248 (2017).
24. Lambertini, L. *et al.* A sensitive functional assay reveals frequent loss of genomic imprinting in human placenta. *Epigenetics Off. J. DNA Methylation Soc.* **3**, 261–269 (2008).
25. Chang, G. *et al.* High-throughput sequencing reveals the disruption of methylation of imprinted gene in induced pluripotent stem cells. *Cell Res.* **24**, 293–306 (2014).
26. Pólvora-Brandão, D. *et al.* Loss of hierarchical imprinting regulation at the Prader–Willi/Angelman syndrome locus in human iPSCs. *Hum. Mol. Genet.* **27**, 3999–4011 (2018).
27. Pastor, W. A. *et al.* Naive Human Pluripotent Cells Feature a Methylation Landscape Devoid of Blastocyst or Germline Memory. *Cell Stem Cell* **18**, 323–329 (2016).
28. Pastor, W. A. *et al.* Naïve human pluripotent cells feature a methylation landscape devoid of blastocyst or germline memory. *Cell Stem Cell* **18**, 323–329 (2016).
29. Yagi, M. *et al.* Derivation of ground-state female ES cells maintaining gamete-derived DNA methylation. *Nature* **548**, 224–227 (2017).
30. Deng, Q., Ramsköld, D., Reinius, B. & Sandberg, R. Single-Cell RNA-Seq Reveals Dynamic, Random Monoallelic Gene Expression in Mammalian Cells. *Science* **343**, 193–196 (2014).

31. Garieri, M. *et al.* Extensive cellular heterogeneity of X inactivation revealed by single-cell allele-specific expression in human fibroblasts. *Proc. Natl. Acad. Sci. U. S. A.* **115**, 13015–13020 (2018).
32. Kim, J., Bretz, C. L. & Lee, S. Epigenetic instability of imprinted genes in human cancers. *Nucleic Acids Res.* **43**, 10689–10699 (2015).
33. Chung, W. *et al.* Single-cell RNA-seq enables comprehensive tumour and immune cell profiling in primary breast cancer. *Nat. Commun.* **8**, 1–12 (2017).
34. Inoue, A., Jiang, L., Lu, F., Suzuki, T. & Zhang, Y. Maternal H3K27me3 controls DNA methylation-independent imprinting. *Nature* **547**, 419–424 (2017).
35. Kolodziejczyk, A. A. *et al.* Single Cell RNA-Sequencing of Pluripotent States Unlocks Modular Transcriptional Variation. *Cell Stem Cell* **17**, 471–485 (2015).

Acknowledgements

The authors thank members of the Martello laboratory for discussions and suggestions. G.M.'s Laboratory is supported by grants from the Giovanni Armenise–Harvard Foundation, the Telethon Foundation (TCP13013) and an ERC Starting Grant (MetEpiStem), C.R. is supported by the Italian Association for Cancer Research (AIRC) [IG21837], P.M. has been supported by the European Molecular Biology Organization (EMBO) Short-Term Fellowship [8517].

Contributions

G.M., C.R. and P.M. designed the study; P.M. and G.S. developed the software and all custom code; P.M. performed all analyses; V.P. and L.D. performed validation experiments; P.M. and L.D. prepared the figures; P.M., C.R. and G.M. wrote the manuscript with input from all authors; G.M. and C.R. supervised the study and provided fundings.

FIGURE 1

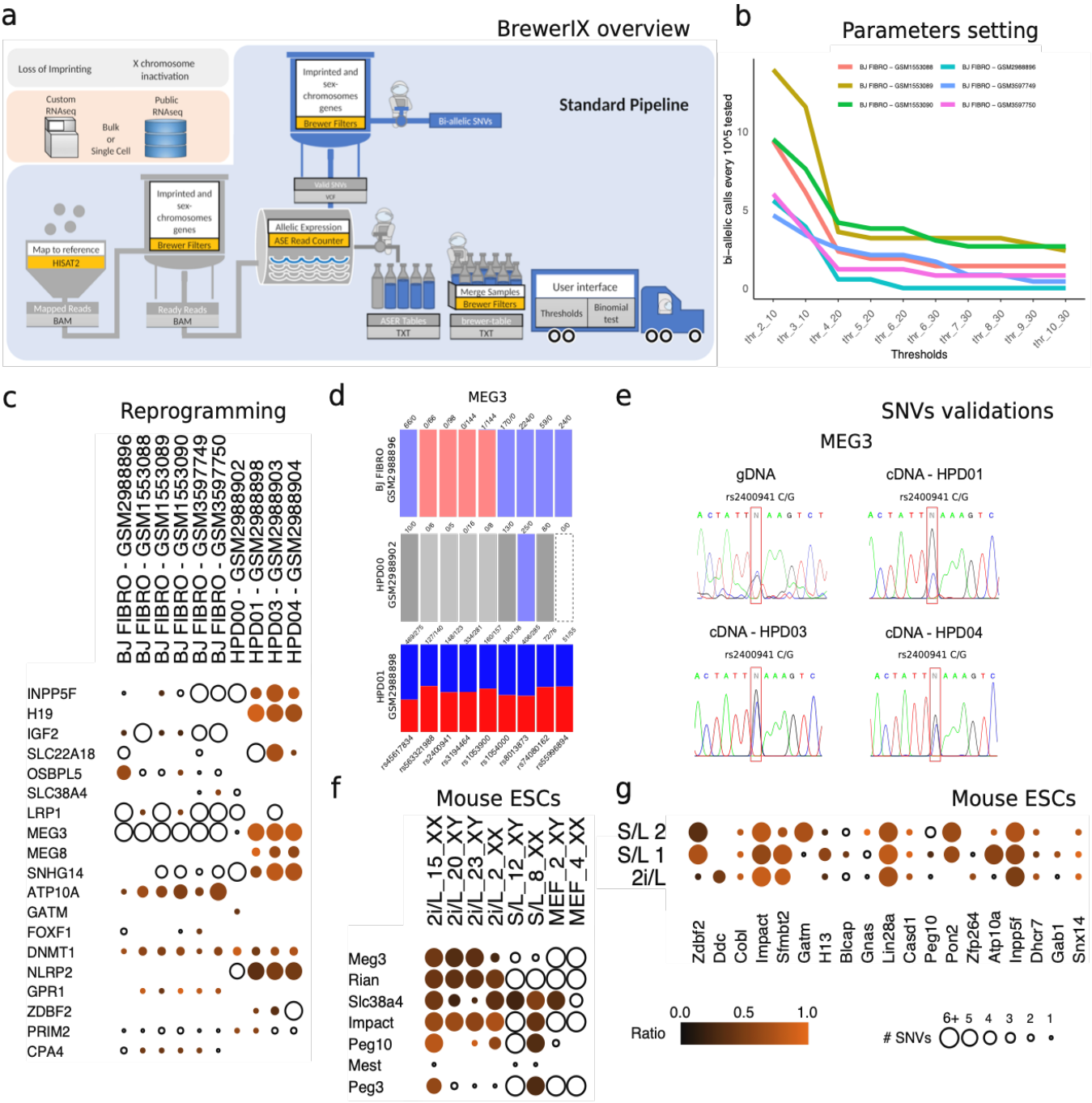

Figure 1. Analyses of Imprinted gene expression in naive pluripotent cells with BrewerIX

a, BrewerIX rational and overall implementation scheme for the Standard pipeline. **b**, False positives bi-allelic calls estimated by analysis of transcripts on the X chromosome in 6 male BJ fibroblasts samples. On the x axis thresholds combination of overall depth and minimal coverage of the minor allele. **c**, BrewerIX gene summary panel results on bulk RNAseq data from isogenic human fibroblasts and primed (HPD00) and naive (HPD01/3/4) iPSCs. The larger the dot, the higher the number of SNVs supporting the bi-allelic call. The brighter the color, the closer to 1 is the average of the allelic ratios (minor/major) of all bi-allelic SNVs. When both alleles are expressed at the same level the allelic ratio is equal to 1. Empty dots indicate detected genes with no evidence of bi-allelic expression, while white dots indicate undetected genes. **d**, BrewerIX SNV summary panel for Meg3 in the case study shown in c. A barplot for each sample is reported with as many bars as the number of SNVs per gene. Solid colors represent actual SNV with both loci expressed, blue and red are the reference and the alternative/minor allele. Transparent colors indicate SNVs detected with no evidence of bi-allelic expression, while grey-scale colors indicate SNVs that do not meet the minimal coverage. **e**, Experimental validation of the indicated SNVs by PCR followed by Sanger sequencing. The SNVs of interest are highlighted by a red box. See Supplementary Table 4 for a list of all SNVs validated. Each SNVs was detected in two independent experiments, using either Forward or Reverse sequencing primers. **f**, BrewerIX gene summary panel results on bulk RNAseq data generated by Yagi and colleagues[29]. Murine ESCs were expanded in either 2i/L or S/L conditions, while mouse embryonic fibroblasts (MEF) serve as controls. **g**, BrewerIX gene summary panel results from bulk RNAseq data of mESCs cultured in 2i/L or S/L (two biological replicates) by Kolodziejczyk and colleagues[35]. See Fig. 2a for matching single-cell RNAseq samples.

FIGURE 2

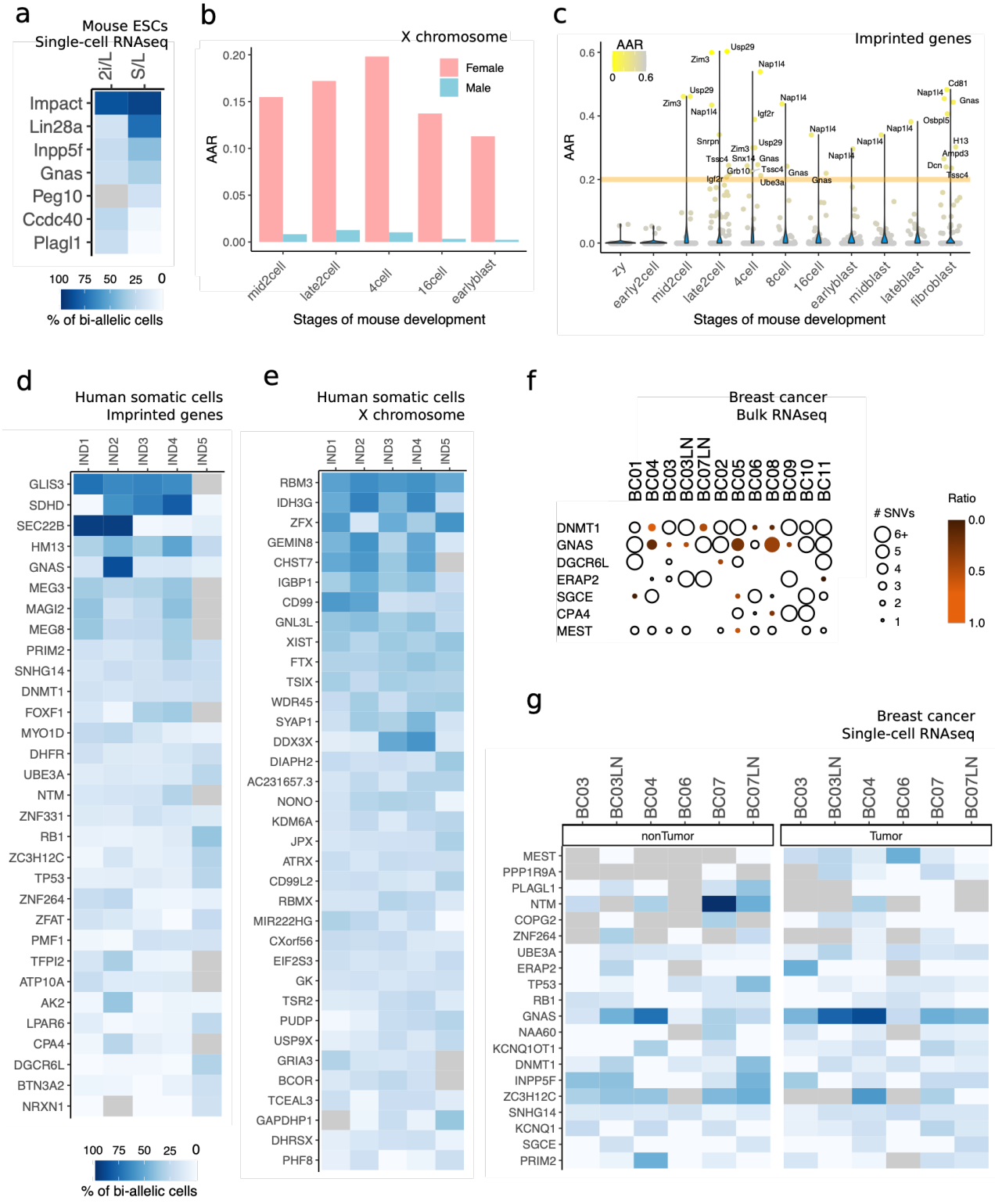

Figure 2. Analyses of single-cell RNAseq data of mouse embryonic and human adult cells

a, Analysis of single-cell RNAseq data from mESCs cultured in 2i/L or S/L, matching those shown in Fig. 1g. Results are summarised as percentages (degree of blue) of cells in which a given gene was expressed bi-allelically. Gray indicates undetected genes. Number of cells analyzed: 2i/L 384, S/L 288. **b**, Average allelic ratio (AAR) defined as the average of minor/major ratios across single cells for all genes in chromosome X in male and female embryonic cells detected by single-cell RNAseq[30]. Number of cells for male (M) and female (F) for each developmental stage: mid2cell 6M, 6F; late2cell 4M, 6F; 4cell 3M, 11F; 16cell 27M, 23F; earlyblast 28M, 15F. See also Supplementary Fig. 6. **c**, Distribution of AAR for imprinted genes across mouse developmental stages. Genes with AAR >= 0.2 are labelled. Number of cells for developmental stage: zygote 4, early2cell 8, mid2cell 12, late2cell 10, 4cell 14, 8cell 47, 16cell 58, earlyblast 43, midblast 60, lateblast 30, fibroblast 10. See also Supplementary Fig. 7. **d**, Analysis of single-cell RNAseq data[16] from 772 human fibroblasts and 48 lymphoblastoid cells from 5 female individuals (IND1-5). Results are summarised as percentages (degree of blue) of cells in which a given gene was expressed bi-allelically. Gray indicates undetected genes. Number cells: IND1 229, IND2 159, IND3 192, IND4 192 and IND5 48. **e**, Results of X chromosome genes on samples described in d. **f**, BrewerIX gene summary panel results from bulk RNAseq data from human breast cancer samples. LN indicates matching metastatic lymph nodes. **g**, Analysis of single-cell RNAseq data from breast cancer samples, matching those analyzed in f. Number of cells: BC03_nonTumor 18, BC03_Tumor 15, BC03LN_nonTumor 43, BC03LN_Tumor 10, BC04_nonTumor 8, BC04_Tumor 47, BC06_nonTumor 10, BC06_Tumor 8, BC07_nonTumor 24, BC07_Tumor 26, BC07LN_nonTumor 26, BC07LN_Tumor 26.