Accurate estimation of single cell allele-specific gene expression using all reads and combining information across cells

Kwangbom Choi[1], **Narayanan Raghupathy**[1], **Gary A. Churchill**[1,*]
[1]The Jackson Laboratory, 600 Main Street, Bar Harbor, Maine, 04609

Abstract. Single-cell RNA sequencing (scRNA-Seq) can reveal features of cellular gene expression that cannot be observed in whole-tissue analysis. Allele-specific expression in single cells can provide an even richer picture of the stochastic and dynamic features of gene expression. Single-cell technologies are moving toward sequencing larger numbers of cells with low depth of coverage per cell. Low coverage results in increased sampling variability and frequent occurrence of zero counts for genes that are expressed at low levels or that are dynamically expressed in short bursts. The problems associated with low coverage are exacerbated in allele-specific analysis by the almost universal practice of discarding reads that cannot be unambiguously aligned to one allele of one gene (multi-reads). We demonstrate that discarding multi-reads leads to higher variability in estimates of allelic proportions, an increased frequency of sampling zeros, and can lead to spurious findings of dynamic and monoallelic gene expression. We propose a weighted-allocation method of counting reads that substantially improves estimation of allelic proportions and reduces spurious zeros in the allele-specific read counts. We further demonstrate that combining information across cells using a hierarchical mixture model reduces sampling variability without sacrificing cell-to-cell heterogeneity. We applied our approach to track changes in the allele-specific expression patterns of cells sampled over a developmental time course. We implemented these methods in extensible open-source software `scBASE`, which is available at `https://github.com/churchill-lab/scBASE`.

Keywords: allele-specific expression, single-cell RNA-Seq, hierarchical mixture model, partial pooling.

* Gary.Churchill@jax.org

INTRODUCTION

In diploid cells, the two allelic copies of a gene can differ in their timing and level of expression due to genetic, environmental, and stochastic factors. Direct sequencing of RNA from whole tissue samples using high-throughput sequencing technologies can provide quantitative information about the abundance and the allelic origin of expressed transcripts. Allelic imbalance is common across many genes [Crowley et al., 2015] and can range from a subtle imbalance to complete monoallelic expression. Analysis of allelic expression in whole tissues can provide insights into the regulation of gene expression [Pastinen, 2010] but it is limited to average expression across many cells and thus cannot reveal features of cell-to-cell heterogeneity in gene expression.

The development of single-cell RNA-Seq (scRNA-Seq) has opened a new perspective on the dynamics of gene expression, revealing details that are obscured in whole tissue analysis. At the single-cell level, timing of the allele-specific transcription of genes is stochastic, resulting in variation in allelic proportions across cells. Single cell analysis has revealed that the expression of many genes occurs in bursts over short time intervals (transcriptional bursting) [Chong et al., 2014, Reinius and Sandberg, 2015]. Transcriptional bursts can occur asynchronously between the maternal and paternal alleles of a gene and, as a result, individual cells may show random monoallelic expression [Reinius and Sandberg, 2015]. The frequency of the different allelic expression states across cells can provide quantitative estimates of parameters that describe the dynamics of gene expression [Jiang et al., 2017].

The analysis of scRNA-Seq data poses challenges due to the limited sampling of the RNA pool in individual cells. Current trends in scRNA-Seq experiments are moving toward lower sequencing depth across larger numbers of cells. While this is a promising strategy to maximize information at a fixed cost, low sequencing coverage increases statistical sampling variation and can result in the frequent observation of zero reads for one or both alleles of a gene. These zero counts can be misinterpreted as monoallelic expression or absence of expression even though the transcripts may be present in a cell. Thus, analysis methods that account for sampling variation are needed to accurately interpret the dynamics of gene expression. The challenges associated with low read coverage and sampling variation are exacerbated by the near-universal practice of discarding multi-mapping reads (or multi-reads) that do not align uniquely to one allele of one gene.

We propose a new method for the analysis of allele-specific expression from scRNA-Seq data, which is implemented in our `scBASE` (single-cell Bayesian analysis of Allele-Specific Expression) software. We first demonstrate the importance of retaining all of the available sequence reads in

single-cell analysis. We then describe a mixture model for allelic proportions based on a probabilistic classification of allelic expression states of each gene in each cell. Bayesian analysis of the mixture model achieves partial pooling of information across cells. Partial pooling provides cell-specific estimation of allelic proportions for each gene and thus it preserves cell-to-cell heterogeneity. It also combines data across cells in similar allelic expression states to improve the precision of estimation as compared to estimation methods that do not share data across cells. We apply the `scBASE` algorithm to published data from a developmental time course of F1 hybrid mouse embryos [Deng et al., 2014] and highlight how allele-specific analysis can reveal dynamic features of gene expression during early embryonic development.

METHODS

Overview of the scBASE model

The `scBASE` algorithm is composed of three steps: *read counting*, *classification*, and *estimation* (Figure 1). The read counting step is applied first to resolve read mapping ambiguity due to multi-reads and to estimate expected read counts. The classification and estimation steps are executed iteratively to classify the allelic expression state and to estimate the allelic proportions for each gene in each cell using a hierarchical mixture model. We have implemented `scBASE` as a Monte Carlo Markov chain (MCMC) algorithm, which randomly samples parameter values from their conditional posterior distributions. We have also implemented the classification and estimation steps as an Expectation-Maximization (EM) algorithm that converges to the maximum a posteriori parameter estimates. We provide a brief description of the algorithm here and provide additional details in Supplemental Methods.

***Read counting*:** In order to count all of the available sequence reads for each gene and allele,

we have to resolve the two types of read mapping ambiguity that occur when aligning reads to a diploid genome. Genomic multi-reads align with equal quality to more than one gene. Allelic multi-reads align with equal quality to both alleles of a gene. In `scBASE`, multi-reads are resolved by computing a weighted allocation based on the estimated probability of each alignment. We use an EM algorithm implemented in EMASE software for this step [Raghupathy et al., 2018]. Alternatively, read counting could be performed using similar methods implemented in `RSEM` [Li and Dewey, 2011] or `kallisto` [Bray et al., 2016] software. The estimated maternal read count (x_{gk}) for each gene (g) in each cell (k) is the weighted sum of all reads that align to the maternal allele, where the weights are proportional to the probability of the read alignment. Similarly, the estimated paternal read count (y_{gk}) is the weighted sum of all reads that align to the paternal allele. The total read count is the sum of the allele specific counts ($n_{gk} = x_{gk} + y_{gk}$). A parameter of interest is the allelic proportion p_{gk}. The read counting step provides an initial estimate $\hat{p}_{gk} = x_{gk}/n_{gk}$, which we refer to as the no-pooling estimator because it uses only the data from one cell.

***Classification*:** In the classification step, we estimate the allelic expression state (z_{gk}) for each gene in each cell. The allelic expression state is a latent variable with three possible values $z_{gk} \in \{P, B, M\}$ representing paternal monoallelic, bi-allelic, and maternal monoallelic expression, respectively. Uncertainty about the allelic expression state derives from sampling variation that can produce zero counts for one or both alleles even when the allele-specific transcripts are present in the cell. We account for this uncertainty by computing a probabilistic classification based on a mixture model in which the maternal read counts x_{gk} are drawn from one of three beta-binomial distributions (given n_{gk}) according to the allelic expression state z_{gk}. For a gene in the bi-allelic expression state the maternal allelic proportion is denoted p^B_{gk} and, as suggested by the notation, it may vary from cell to cell following a beta distribution. For a gene in the paternal

monoallelic expression state, the allelic proportion p_g^P follows a beta distribution with a high concentration of mass near zero. Similarly, for a gene in the maternal monoallelic expression state, we model p_g^M using a beta distribution with the concentration of mass near one. We refer to this as a soft zero distribution. The beta distribution parameters for the maternal and paternal states are gene-specific but are constant across cells.

***Estimation*:** The classification step assumes that the hierarchical mixture model parameters are known. In the estimation step, we employ a Monte Carlo Markov Chain (MCMC) algorithm in which model parameters are drawn repeatedly from their conditional posterior distributions. This model describes gene-specific allelic proportions for each cell and thus it has a very large number of parameters. In the scRNA-Seq setting where thousands of genes are measured but low read counts and sampling zeros are prevalent, we may have limited data to support their reliable estimation. Bayesian analysis of the hierarchical model treats parameters as random variables and is well suited for this type of estimation. In this context, the hierarchical model improves the precision of estimation by borrowing information across cells for each gene, giving more weight to cells that are in most similar allelic expression state. This estimation technique is sometimes referred to as *partial pooling*. Specifically, we sample the mixture weights $\left(\pi_{g\bullet}^P, \pi_{g\bullet}^B, \pi_{g\bullet}^M\right)$ and the class-specific allele proportions $\left(p_g^P, p_{gk}^B, p_g^M\right)$; generate classification probabilities $\left(\pi_{gk}^P, \pi_{gk}^B, \pi_{gk}^M\right)$; and then estimate the allelic proportions as a weighted average

$$p_{gk} = \pi_{gk}^P\, p_g^P + \pi_{gk}^B\, p_{gk}^B + \pi_{gk}^M\, p_g^M \tag{1}$$

The average value across many iterations is $\tilde{p}_{gk}$, the partial pooling estimator.

[Figure 1 about here.]

Data

We applied our model to scRNA-Seq data from 286 pre-implantation mouse embryo cells from an F1 hybrid mating between a female CAST/EiJ (CAST) mouse and a male C57BL/6J (B6) cross [Deng et al., 2014]. The cells were sampled along a time course from the zygote and early 2-cell stage through the blastocyst stages of development. We created a diploid transcriptome from CAST- and B6-specific sequences of each annotated transcript (Ensembl Release 78) [Keane et al., 2011] and aligned reads from each cell to obtain allele-specific alignments. In order to ensure that we had sufficient polymorphic sites for allele-specific expression analysis, we restricted attention to 13,006 genes with at least 4 allelic unique reads in at least 32 out of 286 cells. Additional details are provided in Supplemental Methods.

RESULTS

Weighted allocation reduces the rate of false monoallelic expression calls.

In this and the following section we consider only the counting step of the `scBASE` algorithm in order to demonstrate the impact of weighted allocation in comparison to using only the uniquely mapped read counts for determining allelic expression states. The sequence reads from the developmental time course data include 2.5% genomic multi-reads, 59.3% allelic multi-reads, and 23.3% complex multi-reads, which are both genomic and allelic. In a typical scRNA-Seq workflow for allele-specific expression, these reads are discarded leaving only the unique 14.9% of the original sequence reads for analysis.

To evaluate the impact of discarding multi-reads, we estimated p_{gk} using the no-pooling estimator with weighted allocation counts and with the uniquely mapped read counts. We called genes

as monoallelic maternal if $\hat{p}_{gk} > 0.98$ and monoallelic paternal if $\hat{p}_{gk} < 0.02$ for both estimators. The unique reads method generates a higher rate of monoallelic expression calls (Figure 2A), calling on average $\sim$377 more genes with monoallelic expression in each cell. The weighted allocation method has high rates of monoallelic expression calls in the zygote and 2-cell stages of development where large numbers of transcripts are present only from the maternal allele, thus the weighted allocation method is not simply failing to call monoallelic expressed genes.

We observed 1,393 examples of genes where the unique-reads method fails to call bi-allelic expression in more than 25% of cells compared to the weighted allocation, for example, *Mtdh* (Figure 2B). These genes are consistently bi-allelic in many cells according to weighted allocation, but their pattern of allelic expression based on unique reads is consistent with a dynamic bursting model of allelic expression.

[Figure 2 about here.]

To illustrate how discarding multi-reads can generate false monoallelic expression calls, we consider an example using read counts from the gene *Cdk2ap1* in one cell selected from the 16-cell stage (Figure 3). There are 34,565 reads that align to *Cdk2ap1*. Only 408 reads align uniquely and all of these reads derive from the CAST (maternal) allele (Figure 3A). There are no reads that align uniquely to the B6 (paternal) allele. Thus, based on the unique reads, we would conclude that *Cdk2ap1* has maternal monoallelic expression. Next we look at the multi-reads and consider each of the possible alignment patterns to alleles and to other genes that share multi-reads with *Cdk2ap1* (Figure 3B). There are 2,187 allelic multi-reads that align to both alleles of *Cdk2ap1*, 99 reads align to both alleles of Gm12184, and one aligns to both alleles of Trim7. This indicates that the expression of *Cdk2ap1* is much higher (22-fold) than the combined expression of the other gene

family members. In addition, there are 579 genomic multi-reads that align to only the B6 allele of *Cdk2ap1* and to only the B6 allele of Gm12184. Weighted allocation of these reads between *Cdk2ap1* and Gm12184 should assign the majority of these reads to the more highly expressed gene *Cd2kap1*. These multi-read alignment patterns suggest that both alleles of *Cdk2ap1* are expressed but because the B6 allele has greater similarity to other gene family members, the B6 reads from *Cdk2ap1* are not counted by the unique-reads method. The remaining 31,699 multi-reads align to both alleles of *Cdk2ap1* and to both alleles of at least one other gene. The weighted allocation of these reads (Figure 3C) indicates that *Cdk2ap1* has bi-allelic expression with a maternal allele proportion of $\hat{p}_{gk} = 0.466$. Extending these analysis to all genes, we conclude that discarding multi-reads produces hundreds of spurious monoallelic expression calls per cell creating a false impression of dynamic allele expression that is not seen with the weighted allocation method.

[Figure 3 about here.]

The timing of maternal and paternal allelic bursts is coordinated

Allelic expression of genes is thought to occur in short bursts of transcriptional activity [Reinius and Sandberg, 2015]. A fundamental question regarding the temporal regulation of transcriptional bursting is whether the expression of the maternal and paternal alleles is coordinated or if transcription from one allele occurs independently of the transcriptional state of the other allele. Independence between transcription of two alleles has been assumed in previous model-based analyses of allele-specific expression in single cells [Jiang et al., 2017].

In order to evaluate possible coordination in the timing of allele-specific transcriptional bursts we examined the co-occurrence of maternal and paternal gene expression. For each of 286 cells we

classified 13,006 genes into one of four allelic expression states: not expressed, maternal monoallelic, paternal monoallelic, or bi-allelic using a threshold of >2 reads to determine that an allele is expressed for both weighted-allocation and unique reads counts. (We varied this threshold from 0 to 8 with little overall impact on our conclusions.) We then constructed a 2×2 cross classification of maternal and paternal allele expression. Each gene has a 2×2 table that counts the numbers of cells in each of the four categories of allele-specific expression (Figure 4A). We restricted attention to *dynamic* genes that had at least 4 cells in each category of contingency table. The weighted allocation method identified 9,960 dynamic genes and unique reads identified 10,394 dynamic genes (Figure 4B). For each dynamic gene, we evaluated the statistical independence of maternal and paternal allelic expression by computing the log2 of the odds ratio (logOR). Values of the logOR near zero are consistent with independence. A significantly positive logOR indicates correlated allelic expression (coordinated bursts) and negative values indicate anti-correlated expression.

The mean value of weighted allocation logOR is 2.44 and the large majority of all genes displayed significantly positive values of logOR (8,629 out of 9,960 at FDR=0.05), consistent with a strong and almost universal tendency for allelic bursts to occur synchronously (Figure 4C and 4D). Interestingly, 93 genes (e.g., *Padi2*, *Dqx1*, and *Snprn*) showed significant anti-correlation in their expression (at FDR=0.05). The distribution of logOR based on unique read counts is smaller in magnitude with mean logOR of 1.76. The reduced absolute value of logOR for unique reads may have contributed to the impression that transcriptional bursting occurs independently [Jiang et al., 2017].

[Figure 4 about here.]

We concluded that these data do not support the independence of allelic transcriptional bursts

when all the aligned reads are considered. By calling spurious monoallelic expressions, unique-reads method makes allelic bursting look more dynamic.

Allelic expression states evolve over a developmental time course

To demonstrate the utility of classifying the allelic expression states of a gene we applied the Classification and Estimation steps of `scBASE` to the weighted allocation counts of the Deng et al. data. In this section, we focus on the classification of genes and interpretation of the allelic expression states across populations of cells.

For many genes, the allelic proportion is highly variable across cells due to the inherent stochasticity of allelic expression in individual cells. It is useful to distinguish different classes of cells that show monoallelic paternal (P), bi-allelic (B), or monoallelic maternal (M) expression of a given gene. Here we apply a hierarchical mixture model to compute the posterior probability of allelic expression states of genes in each cell. This probabilistic classification allows for uncertainty associated with the statistical sampling from the pool of transcripts that are present in the cell. The uncertainty will be greater when the read coverage is lower or for genes with fewer polymorphisms to distinguish between alleles.

The expected proportions of cells in states P, B, and M can be represented as point in a triangular (ternary) diagram. To interpret the patterns of gene expression across cells, we designate seven patterns of allelic expression (Figure 5 and Supplemental Methods). Genes that are predominantly expressed as P, B, or M will appear near the corresponding vertex of the triangle ($\mathbf{P}$, $\mathbf{B}$ or $\mathbf{M}$ region). Genes with mixed allelic states will appear along the edges ($\mathbf{PB}$, $\mathbf{BM}$, or $\mathbf{MP}$) or near center of the triangle (all three states, $\mathbf{PBM}$). For example, the gene *Pacs2*, which is expressed from either the maternal or the paternal allele but rarely both, is classified as an $\mathbf{MP}$ gene. The

bi-allelic region (**B**) includes genes that may show allelic imbalance ($p_{gk} \neq \frac{1}{2}$) across many cells but are consistently expressed from both alleles e.g., *Mtdh*. The **PB** and **BM** regions include genes that show a mixture of bi-allelic and monoallelic expression with a strong allelic imbalance, e.g., *Timm23* and *Tulp3*. The majority of genes (56.9%) in the blastocyst stages of development are in the **PBM** region (Supplemental Figure S1). These genes display a mix of mono- and bi-allelic expression states (e.g., *Akr1b3*) that is consistent with dynamic allele-specific gene expression with a low bursting rate. Genes that are predominantly bi-allelic may still be expressed in allele specific bursts. However the frequency of transcriptional bursting is high relative to the half-life of the mRNA such that transcripts from both alleles are usually present. The relative timing of allelic bursting may increase cell-to-cell variation in the allelic proportion beyond expectations from a simple binomial sampling model.

[Figure 5 about here.]

Next we examined how allelic expression patterns change over the developmental time course (Figure 6 and Supplemental Figure S1, S2, and S3). In the zygote and early 2-cell stages, essentially all genes show monoallelic maternal expression. At this stage, the hybrid embryo genome is not being transcribed and the mRNA present is derived from the mother (inbred CAST genome). At the mid 2-cell stage the hybrid embryo is being transcribed and we start to see expression of the paternal allele for some genes. Many genes exhibit the **M** and **BM** patterns through the 8- or 16-cell stages perhaps due to the persistence of long-lived mRNA species that were present at the 2-cell stage. The bi-allelic class **B** dominates the late 2-cell and 4-cell stages indicating high levels of expression at rates that exceed the half-life of most mRNA species. In the later stages of development, 8-cell through late blastocyst, most genes transition into the **PBM** pattern. These

genes have rates of bursting that are on the same order of magnitude as the half-life of the mRNA resulting in a stochastic expression pattern at the single cell level. The next most frequent pattern is bi-allelic expression. Genes with purely maternal or paternal allelic expression occur at roughly equal frequency in the blastocyst stages and represent a total of 12.4% of all genes.

[Figure 6 about here.]

Genes with the alternating expression pattern (MP) are rare in these data but there is a set of 200-300 genes that show stable alternating expression throughout the blastocyst stages of development stages (Supplemental Figure S2). One possible explanation is that expression bursting is infrequent relative to the half-life of the transcripts. However, many of these genes have significant negative log-odds ratios indicating that the allelic bursts are not occurring at random; their allelic expression appears to be regulated such that only one allelic copy is expressed at any given time. One possible explanation is that these genes have fixed monoallelic expression in which one allele has been selected (at random) to be expressed and the other allele is suppressed [Reinius and Sandberg, 2015] throughout this period of development. This would imply the existence of a mechanism to ensure that gene-specific monoallelic expression is established and maintained in each cell.

We also identified genes that undergo transitions in their pattern of allelic expression over the development time course (Supplemental Figure S3). Much of the switching between patterns is occurring at the boundaries of the regions in Figure 5 – these boundaries are convenient for interpretation but are somewhat arbitrary. However there are $\sim$400 genes that make dramatic transitions across this space. For example, *Akr1b3* (Figure 7) starts out at the zygote and early 2-cell stage with only maternal alleles present. It transitions to bi-allelic expression by the mid

2-cell stage indicating the onset of transcription of the paternal allele. It then transitions through the paternal monoallelic state. Our interpretation is that the early maternally derived transcripts were present prior to fertilization and these transcripts are still present when the paternal allele in the hybrid embryo gene starts to express. The early maternal transcripts are largely degraded by the 4- to 8-cell stages where we see only expression from the paternal allele. In the early blastocyst stages, we start to see embryonic expression of maternal alleles resulting in a bi-allelic expression pattern by the late blastocyst stage.

[Figure 7 about here.]

Partial pooling of information across cells improves estimation of allelic expression while preserving cell-to-cell heterogeneity

In this section we focus on the estimation of the allelic proportion p_{gk} and carry out an empirical evaluation of several estimation methods. We are most interested in how these estimators performs when read coverage is low to reflect current trends in scRNA-Seq experiments. The data provided by Deng et al. have relatively high coverage of $\sim$13.8M reads per cell across the 60 cells from the mid-blastocyst stage. We down-sampled these data by randomly selecting 1% of reads to obtain an average coverage of $\sim$138k reads per cell. We repeated the sampling 6 times and estimated allelic proportions using each of four methods: (1) the no-pooling estimators based on unique-reads, (2) the no-pooling estimator with weighted allocation, (3) the partial pooling estimator using a simple beta-binomial model, and (4) the 3-component beta-binomial mixture model. We compared the estimated allelic proportions from the down-sampled data to estimates obtained from the full data using the weighted allocation method with no pooling. The latter are based on 100-fold more data and provide an approximate truth standard. The advantage of this approach

is that we are not forced to make detailed modeling assumptions that might bias our evaluations if we relied on simulated data. A similar approach to evaluation of methods single-cell data was employed by [Huang et al., 2018]. Until we achieve a better understanding of the technical and biological factors that determine the behavior of single-cell gene expression, it is not clear how to best simulate the process.

`scBASE` reported partial pooling estimators on 5,265 genes with this simulated data. We first compare the no-pooling estimators based on weighted allocation and unique reads (Figure 8A). We computed differences in the mean squared error (MSE) of estimated allelic proportions between two estimation methods. Points above the reference line (y=0) indicate that the weighted allocation estimator is closer to the truth and points below the lines indicate genes where the unique reads estimator is better. The weighted allocation method improves estimation for most genes (2,592 versus 200 out of 2,793 genes) and the average MSE difference is 0.121. We note that unique-reads method was not able to report allelic proportions on 2,472 genes out of 5,265.

Next we compared the `scBASE` mixture model (partial pooling) to weighted allocation (no pooling) (Figure 8B). The mixture model improves estimation for the majority of genes (4,718 versus 405 out of 5,265) with an average MSE difference of 0.075. The greatest gains are seen in the low expression range (<10 reads per gene). For the most highly expressed genes, there is no reduction in MSE, which is consistent with our expectation that pooling of information across genes is most impactful when coverage is low.

Next we considered a simplified version of the `scBASE` algorithm in which the 3-component mixture model is replaced by a single component beta-binomial model (Figure 8C). The beta-binomial model performs substantially better (in 4,678 versus 445 genes out of 5,265) than the weighted allocation estimator, with an average MSE difference of 0.068. The greatest improve-

ments are seen again for the low expressed genes. Comparison of the single-component model to mixture models (Figure 8D) indicates that the mixture model provides improved estimation (in 3,649 versus 1,489 genes out of 5,265) but the gains are modest (with an average MSE difference of 0.0068). The relatively good performance of the single-component model reflects the adaptability of the beta distribution – the parameters of the single-component beta distribution can be unimodal, heavily skewed toward one, or toward zero, or U-shaped. Therefore a single beta-binomial can represent a variety of distributions of allele proportion for genes with extreme allelic expressions. The advantage of the 3-component mixture model is that it provides an estimate of the allelic expression state of each gene in each cell. As illustrated above, classification of allele expression states can add to our understanding of the dynamics of allele-specific expression.

[Figure 8 about here.]

DISCUSSION

Sequence reads that do not align uniquely to a gene, genomic multi-reads, are often ignored in RNA-Seq analysis. This can lead to less accurate and potentially biased estimates of total gene expression [Degner et al., 2009, Li and Dewey, 2011, Bray et al., 2016, Raghupathy et al., 2018]. In allele-specific RNA-Seq analysis, there is additional ambiguity due to reads that do not distinguish between alleles of a gene. Allelic multi-reads can be abundant and previous reports of allele-specific expression using scRNA-Seq have discarded allelic multi-reads prior to analysis. The counting step of `scBASE` uses an EM algorithm to obtain a weighted allocation of read counts that includes genomic and allelic multi-reads. We demonstrate that this leads to substantial improvement in estimation of allele proportions and that it reduces the occurrence of false monoallelic expression calls.

In whole-tissue RNA-Seq with high read depth and weighted allocation of read counts, using only data from one sample will often yield precise and accurate estimates of allelic proportions. However in scRNA-Seq, the total number of reads available from any single cell may be limited and the number of parameters to estimate is large. We propose a hierarchical Bayesian analysis to estimate gene- and cell-specific allele proportions that can improve estimation by combining data across all of the cells. The hierarchical model assumes that the model parameters represent random draws from prior distributions and we use data across all of the cells to estimate these prior distributions. This is how the *partial pooling* of information across cells is achieved. The partial pooling estimates of allelic proportions vary from cell to cell but the variation is moderated relative to the no-pooling estimator $\hat{p}_{gk}$, which uses data from only one cell. With partial pooling information is shared among cells when the gene is likely to be in the same allelic expression state and less sharing occurs between cells that are likely to be in different allelic expressions states. Genes with high read counts are less influenced by partial pooling and the estimated allelic proportions will be close to the no-pooling estimate. Genes with low read counts are most influenced by partial pooling and these show the greatest improvements in overall precision and accuracy.

In the `scBASE` model, monoallelic expression are called by classification using a mixture model. The allelic proportions within each class are bounded away from zero and one ($0 < p^s_{gk} < 1$) and thus there is always some, perhaps very small, possibility of observing a read from either allele. For the monoallelic expression classes we use prior distributions that constrain p^P_g and p^M_g to be near zero or one, respectively. Whenever a gene has multi-reads, the weighted allocation algorithm may produce a fractional but non-zero count for each multi-read. While these values are usually small, they are not exactly zero thus we need the soft-zero model to avoid classifying all such genes as bi-allelic. In addition the soft-zero provides robustness to sequencing errors. To estimate p^P_g and

p_g^M, we use complete pooling across cells. Each gene will have its own rates for the occurrence of reads from the non-expressed allele but there is no cell-to-cell variation in these rates. The allelic proportion for the bi-allelic class, p_{gk}^B, is allowed to vary from cell to cell in the `scBASE` model and estimation uses partial pooling.

The majority of genes display significant positive correlation in the expression of maternal and paternal alleles within a cell. The likely explanation for this coordination is the presence (or absence) of diffusible transcriptional regulators that activate (or suppress) expression and have equal access to both allelic copies of a gene. We identified ∼93 genes that show significant negative correlation. These genes are found in the **MP** class across the later developmental stages (Figure 6 and Supplemental Figure S1). They display monoallelic expression in most cells but in any particular cell either allele may be expressed. These genes could be examples of fixed random monoallelic expression model (*Fixed aRME*; [Reinius and Sandberg, 2015]) in which one randomly selected allele is expressed throughout the developmental time course. Previous studies have presented evidence to support the independent expression of alleles [Jiang et al., 2017] but their analysis was based on unique reads, which may have reduced the apparent strength of correlations.

The `scBASE` model does not include a class that represents the absence of expression. We estimate allelic proportions conditional on the total read counts. In the case where a gene has a total read count of zero in a cell ($n_{gk} = 0$), the classification probabilities reflect the population average. In the probabilistic classification, cells with more reads are classified with greater precision and contribute more to the partial pooling estimates. Cells in which the read count of a gene is zero do not impact the estimation or classification of the gene in other cells.

In the example data, the cells were obtained across multiple developmental stages and they can be grouped according to these stages. In other experiments, the cell types may not be known in

advance but groups of cell may be defined, e.g., by clustering. The scBASE model should be fit to all of the cells in an experiment as a whole. This increases the opportunities to observe genes in each of the possible allelic expression states. For example, if we fit scBASE to only cells in the zygote and 2-cell developmental stages that are dominated by maternal gene expression, we would have little data to estimate the paternal gene expression parameters. When the distribution of the allelic states across a specific group of cells is of interest, we can obtain averages across only those cells as we did for the developmental time course data (see Figure 6 and Figure 7). Specifically, we sum the classification probabilities π_{gk} across the index k corresponding to cells of the desired type.

The MCMC version of the scBASE algorithm and its implementation in the STAN programming language [Carpenter et al., 2017] are easy to modify. The distributions that we assumed for the hierarchical model can be substituted with alternative choices. In the current implementation of scBASE, the read-counting step is carried out separately from the classification and estimation steps. We have implemented an extension of scBASE that incorporates the read counting step into the MCMC iterations. We have also implemented a version of scBASE that incorporates partial pooling across genes. Both of these extensions proved to be impractically slow and did not produce substantially different results on small datasets when compared to the version of scBASE presented here. The scBASE algorithm achieves substantial improvements over current methods for allele-specific scRNA-Seq analysis by including data from multi-reads using weighted allocation and by partial pooling of information across cells. These are general principles that could be applied to other inference problems with scRNA-Seq data.

References

[Bray et al., 2016] Bray, N. L., Pimentel, H., Melsted, P., and Pachter, L., 2016. Near-optimal probabilistic rna-seq quantification. *Nature Biotechnology*, **34**(5):525–527.

[Carpenter et al., 2017] Carpenter, B., Gelman, A., Hoffman, M., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M., Guo, J., Li, P., and Riddell, A., *et al.*, 2017. Stan: A probabilistic programming language. *Journal of Statistical Software, Articles*, **76**(1):1–32.

[Chong et al., 2014] Chong, S., Chen, C., Ge, H., and Xie, X., 2014. Mechanism of transcriptional bursting in bacteria. *Cell*, **158**(2):314–326.

[Crowley et al., 2015] Crowley, J. J., Zhabotynsky, V., Sun, W., Huang, S., Pakatci, I. K., Kim, Y., Wang, J. R., Morgan, A. P., Calaway, J. D., Aylor, D. L., *et al.*, 2015. Analyses of allele-specific gene expression in highly divergent mouse crosses identifies pervasive allelic imbalance. *Nat Genet*, **47**(4):353–360.

[Degner et al., 2009] 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**(24):3207–3212.

[Deng et al., 2014] 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**(6167):193–196.

[Huang et al., 2018] Huang, M., Wang, J., Torre, E., Dueck, H., Shaffer, S., Bonasio, R., Murray, J. I., Raj, A., Li, M., and Zhang, N. R., *et al.*, 2018. Saver: gene expression recovery for single-cell rna sequencing. *Nature Methods*, **15**:539–542.

[Jiang et al., 2017] Jiang, Y., Zhang, N. R., and Li, M., 2017. Scale: modeling allele-specific gene expression by single-cell rna sequencing. *Genome biology*, **18**(1):74.

[Keane et al., 2011] Keane, T. M., Goodstadt, L., Danecek, P., White, M. A., Wong, K., Yalcin, B., Heger, A., Agam, A., Slater, G., Goodson, M., *et al.*, 2011. Mouse genomic variation and its effect on phenotypes and gene regulation. *Nature*, **477**(7364):289–294.

[Li and Dewey, 2011] Li, B. and Dewey, C. N., 2011. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics*, **12**:323.

[Pastinen, 2010] Pastinen, T., 2010. Genome-wide allele-specific analysis: insights into regulatory variation. *Nat Rev Genet*, **11**:533–538.

[Raghupathy et al., 2018] Raghupathy, N., Choi, K., Vincent, M. J., Beane, G. L., Sheppard, K. S., Munger, S. C., Korstanje, R., Pardo-Manual de Villena, F., and Churchill, G. A., 2018. Hierarchical analysis of rna-seq reads improves the accuracy of allele-specific expression. *Bioinformatics*, **34**(13):2177–2184.

[Reinius and Sandberg, 2015] Reinius, B. and Sandberg, R., 2015. Random monoallelic expression of autosomal genes: stochastic transcription and allele-level regulation. *Nat Rev Genet*, **16**(11):653–664. Review.

Figure Captions

Figure 1: Overview of the scBASE algorithm. We summarize the three steps of the `scBASE` algorithm. The Counting Step estimates the expected read counts using an EM algorithm to compute a weighted allocation of multi-reads. Each read is represented as an incidence matrix that summarizes all best-quality alignments to genes and alleles ①. Weighted allocation of multi-reads uses a current estimate of allele-specific gene expression to compute weights equal to the probability of each possible alignment ②. The weights are summed across reads to obtain the expected read counts for each gene and allele ③. Steps ② and ③ are repeated until the read counts converge. The no-pooling estimates of maternal allelic proportion ($\hat{p}_{gk}$) are obtained at this step. The Classification Step computes the posterior probability of paternal monoallelic (P), bi-allelic (B), or maternal monoallelic (M) expression (π^s_{gk}) using current estimates of the model parameters (Equation 3 in Supplemental Methods). The classification model is a beta-binomial mixture model with three components. Model parameters define the weights of mixture components ($\pi^s_{g\cdot}$) and parameters of the Beta densities (α^s_g, β^s_g) that define the distribution of the within-class the maternal allelic proportions (p^s_g). The distributions for monoallelic classes (M and P) define gene-specific soft zeros/ones. The Estimation Step re-estimates the model parameters (in the EM version of `scBASE`) or re-samples parameter values from their posterior distribution (in the MCMC version of `scBASE`). The partial pooling estimate of the maternal allelic proportions ($\tilde{p}_{gk}$) is obtained as an average of the class-specific proportions weighted by the class membership probabilities (Equation 1). The Classification and Estimation steps are iterated until the parameter estimates converge.

Figure 2: Retaining multi-reads reduces monoallelic expression calls. For each of 13,006 genes (see Methods), we obtained the allele-specific read counts using only unique reads and using the weighted allocation algorithm. We counted the numbers of genes in each cell that showed either maternal or paternal monoallelic expression and display the results as points (one per cell) overlaid on boxplots. The zygote and 2-cell stage cells (highlighted in red) have large numbers of genes with maternal monoallelic expression. The outlier cell with high levels of paternal monoallelic expression was noted in Deng et al. On average there are $\sim$377 fewer monoallelic calls per cell with the weighted allocation counts (A). We selected one gene (*Mtdh*) to illustrate the distribution of maternal (X-axis) and paternal (Y-axis) counts across 286 cells. The weighted allocation counts (green) are connected to their corresponding unique counts by a line in the scatter plot. The unique counts resulted in 88 cells with monoallelic expression while only 6 monoallelic calls were seen with weighted allocation (See Figure 4A) (B).

Figure 3: An illustration of how weighted allocation corrects false monoallelic calls. We examined the read count data for *Cdk2ap1* in a single cell from the 16-cell stage of development. There are 408 unique reads and all of them align to the CAST (maternal) allele of *Cdk2ap1*, which is consistent with monoallelic expression of this gene (A). We enumerated the reads that fall into each of the possible multi-alignment patterns (B). Among the allelic multi-reads, there are 2,187 reads that align equally well to both alleles of *Cdk2ap1* and 100 reads aligned uniquely to both alleles of another gene family member. There are 579 genomic multi-reads and all of these align to the B6 allele of *Cdk2ap1* and at least one other gene. The remaining 31,699 reads align equally well to both alleles of *Cdk2ap1* and at least one other gene family member. The allelic and genomic multi-reads provide information needed to estimate the expected distribution of alignments

for all of the reads (C). The weighted allocation algorithm predicts that *Cdk2ap1* in expressed as a bi-allelic gene with a maternal allelic proportion $\hat{p}_{gk} = 0.466$.

Figure 4: Coordinated expression of maternal and paternal alleles. We computed $\log_2$ odds ratio (logOR) for the cross-classification of maternal and paternal expression across 286 cells. For each gene, we built a pair of 2×2 contingency tables: one with class cell counts by weighted allocation, and the other one by unique-reads method. The tables for the gene *Mtdh* are shown for example (A). We computed logOR for the genes with at least 4 cells in every class so that we only see genes that are "dynamic" with respect to allelic expression state. There were 9,960 (=9,131+829) dynamic genes out of 13,006 according to weighted allocation method (B). Density plots of logOR are shown for weighted allocation, unique-reads, and for gene simulated under the assumption of independent activation of maternal and paternal allelic expression. Values of logOR near zero are consistent with independent allelic expression; values greater than zero indicate higher than chance frequency of co-expression; and values less than zero indicate a non-random alternating pattern of allelic expression. The majority of genes show evidence for coordinated expression of positive correlation (C). A density scatterplot of logOR by unique-reads method (Y-axis) against logOR by weighted allocation (X-axis) shows that the latter provides higher values of logOR. In addition, weighted allocation is more robust than unique-reads method as it is based on more data that include genomic and allelic multi-reads. A small number of genes (e.g. *Pacs2*) have significant negative values of logOR. This is consistent with a model in which the expression of one allele exclude the expression of the other (D).

Figure 5: Classification of allele-specific expression patterns across cells. For each gene in each cell we estimate the classification probability π^s_{gk}, where s indicates paternal monoallelic (P), bi-allelic (B), or maternal monoallelic (M) expression. We can then obtain the average proportion of each category across any group of cells ($\pi^s_{g\cdot}$) and represent the expression pattern of the gene as a point in a triangular simplex diagram. A gene that is predominantly paternal, bi-allelic, or maternal expressed across the cell population will be plotted near the corresponding vertex. Points representing genes with mixed classification states across the cell population will appear along the edges or in the center of the triangle. We delineate seven types of gene behavior as indicated by the different colored regions in the diagram: $\mathbf{P}$ (blue), $\mathbf{B}$ (yellow), $\mathbf{M}$ (red), $\mathbf{PB}$ (green), $\mathbf{BM}$ (orange), $\mathbf{MP}$ (purple), and $\mathbf{PBM}$ (gray). Examples of genes from each region are shown for the blastocyst cell population (n=286) around the outside of the diagram. For example, the gene *Pacs2* is expressed from either the maternal or the paternal allele but rarely both and is classified as an $\mathbf{MP}$ gene. The bi-allelic region ($\mathbf{B}$) includes genes that may show allelic imbalance ($p_{gk} \neq \frac{1}{2}$) across many cells but consistently express both alleles (e.g., *Mtdh*). The $\mathbf{PB}$ and $\mathbf{BM}$ regions will include genes that show a mixture of bi-allelic expression and monoallelic expression. Many of the genes in these regions have allelic imbalance for cells in the bi-allelic state and the monoallelic cells could reflect sampling zeros in the lower expressed allele (e.g., *Tmim23* and *Tulp3*). The majority of genes (56.9%) in the blastocyst cells are in the $\mathbf{PBM}$ region and they display a mix of mono- and bi-allelic expression states (e.g., *Akr1b3*).

Figure 6: The distribution of allele-specific expression states changes across developmental stages. Cells were collected across nine developmental stages as indicated on the X-axis. The cell types and numbers of expressed genes at each stage are indicated in parentheses on the X-axis.

For each stage, we counted the proportion of expressed genes that fall into each of the seven allelic expression patterns (Y-axis), indicated by lines using the same color coding used in Figure 5. In the zygote and early 2-cell stage, most genes show purely maternal expression (**M**). The proportion of maternally expressed genes decreases through subsequent stages of development. The numbers of genes showing purely paternal expression (**P**) is low across all developmental stages. The **M** and **P** classes become equally represented in the later stages of development. The 2- and 4-cell stages show high levels of bi-allelic expression (**B**) and the mixed class (**PBM**) proportion becomes highest by the 8-cell stage.

Figure 7: Allele-specific expression pattern of Akr1b3 changes over the developmental time course. The proportions of allele-specific expression classes of *Akr1b3* over the developmental time course are shown as a trajectory of line segments in the simplex (yellow to blue color gradient indicates transitions between stages) (A). Scatterplots of maternal versus paternal read counts of *Akr1b3* are shown at each developmental stage. Each point represents one cell. This gene starts in the maternal monoallelic state (**M**), it transitions through **PBM** to a paternal expression state (**P**), and then transitions to bi-allelic expression (**B**) in the blastocyst stages (B).

Figure 8: The `scBASE` model improves allele-specific estimates when total read coverage is low. A comparison between before and after partial pooling on simulations using 1% randomly sampled reads from the original full data (60 cells in mid-blastocyst stage). This simulation data set has 138,193 reads in average. We repeated this simulation 6 times to make sure if the trend persists. X-axis is the expression level (in expected read counts) of the simulated data set, and Y-axis is the difference between deviations (in MSE) from the ASE results using original full depth data. We benchmarked the weighted allocation versus unique-reads method (A), `scBASE` 3-component Beta-Binomial mixture model versus weighted allocation (B), single-component Beta-Binomial model versus weighted allocation (C), and `scBASE` 3-component Beta-Binomial mixture model versus single-component Beta-Binomial model (D). After partial pooling with the hierarchical models – single-component or 3-component mixture Beta-Binomial, the accuracy improved on 4,678 and 4,718 genes (y>0) out of 5,265 genes respectively whereas the accuracy decreased on 445 and 405 genes (y<0). The heatmap clearly shows that our mixture model rescues low expressed genes especially ones with expression level of <10 reads. The improvement is higher as the depth of coverage gets lower, which we can see the cluster above y>0 is larger. But the improvement by 3-component Beta-Binomial mixture model was modest compared to single-component Beta-Binomial model, arguably because of the flexibility of Beta distribution.

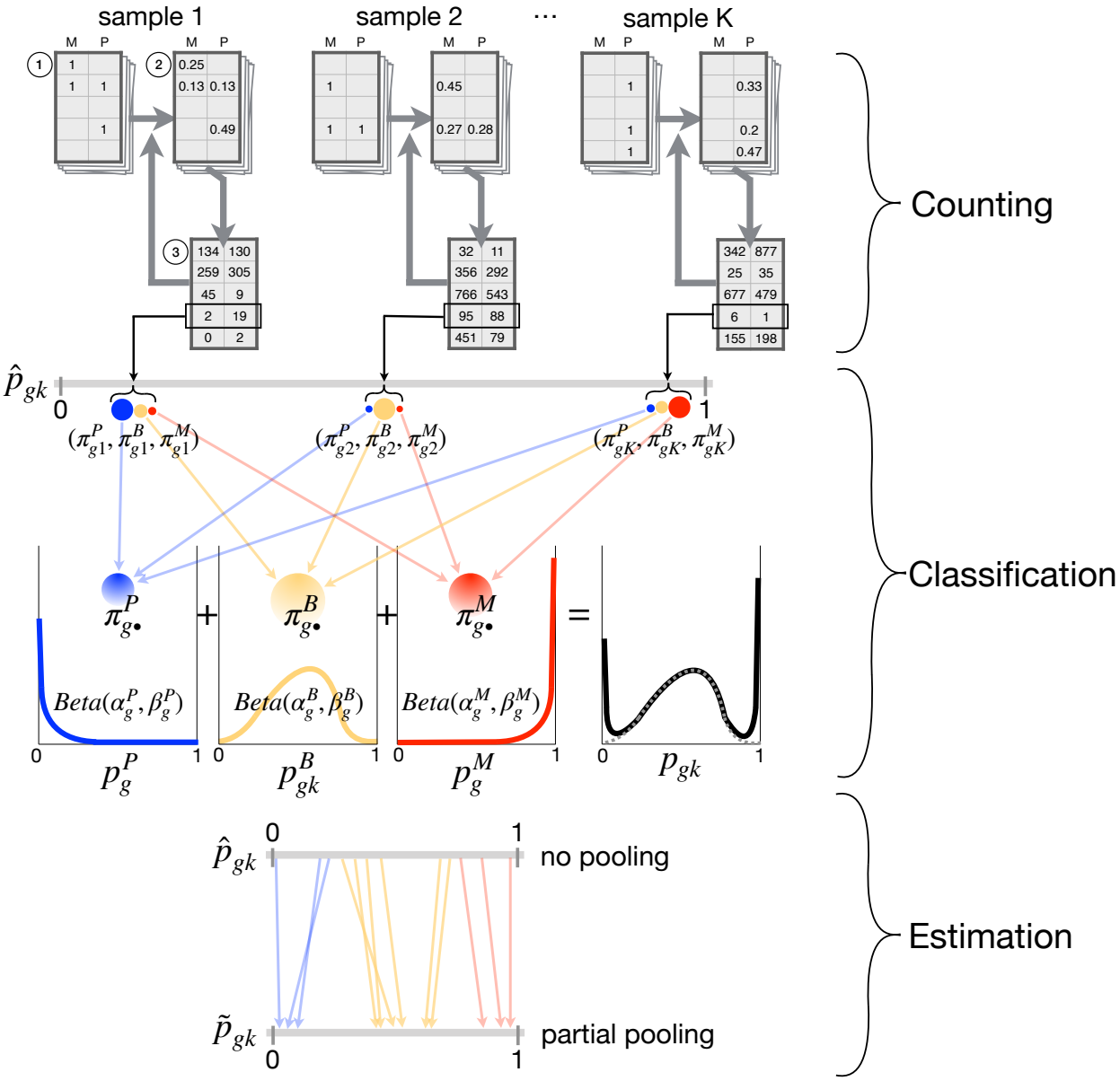
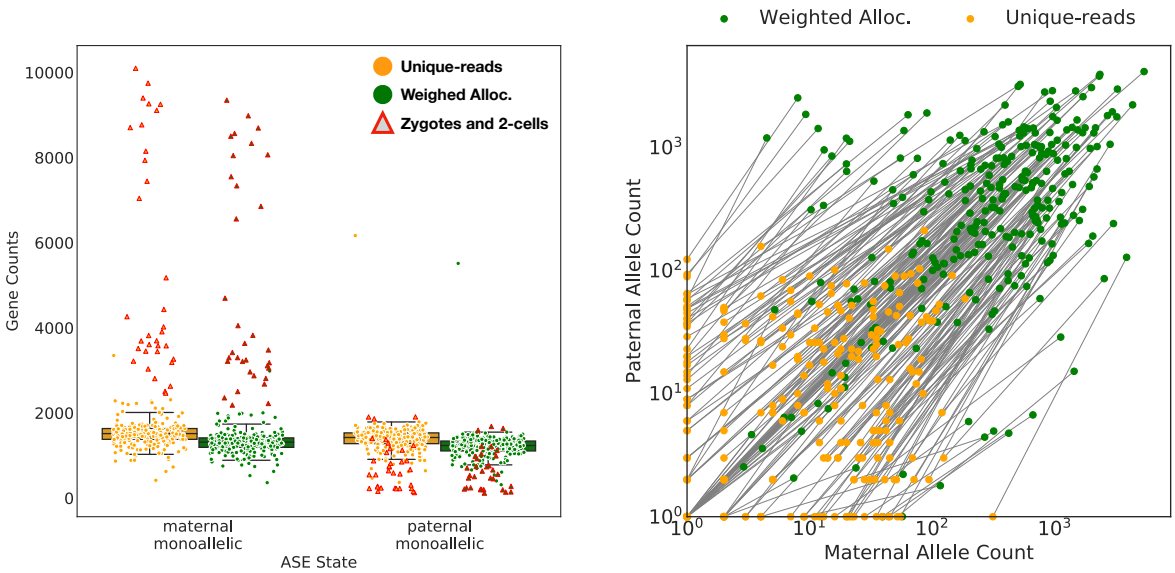


Figure 1



(A) (B)

Figure 2

(A)

	CAST/EiJ	C57BL/6J
Cdk2ap1	408	0
Others	0	0

(C)

	CAST/EiJ	C57BL/6J
Cdk2ap1	14350.3	16421.8
Others	652.4	652.4

(B)

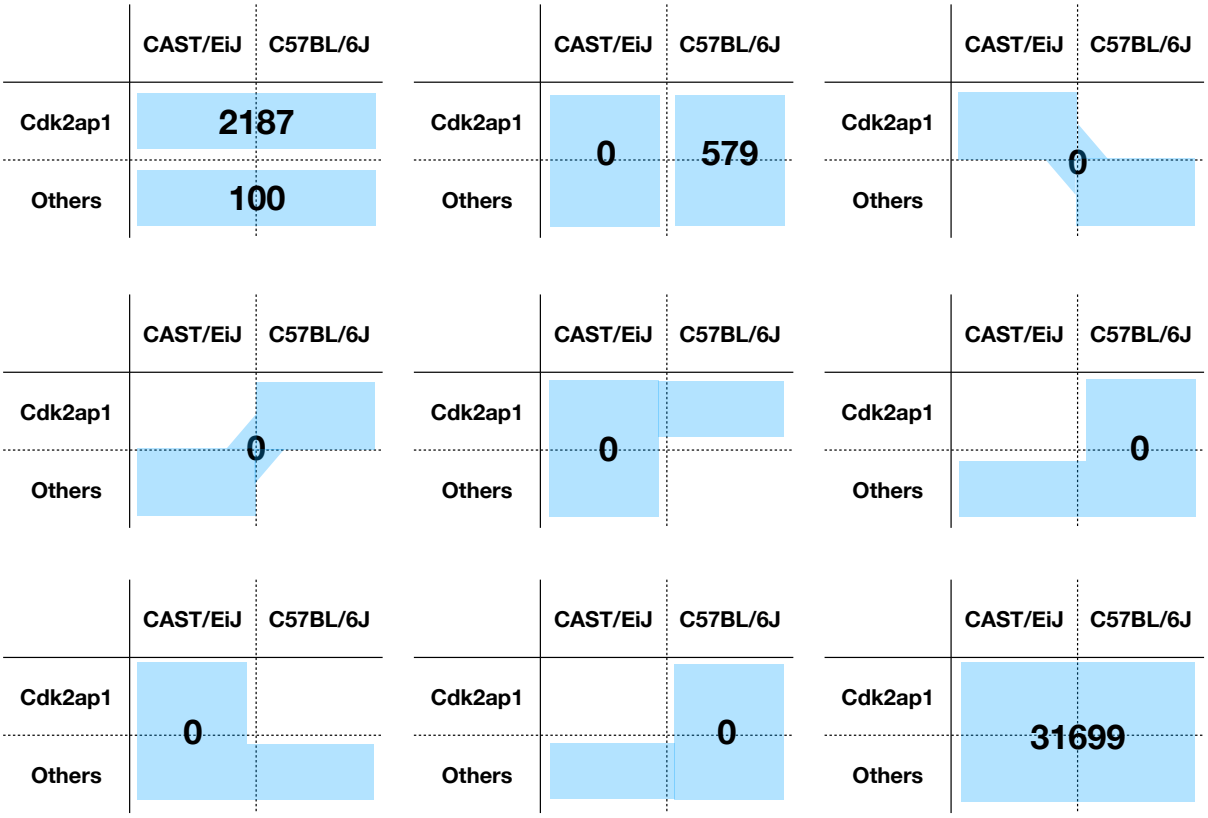
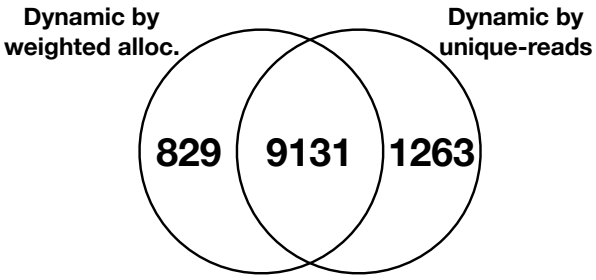


Figure 3

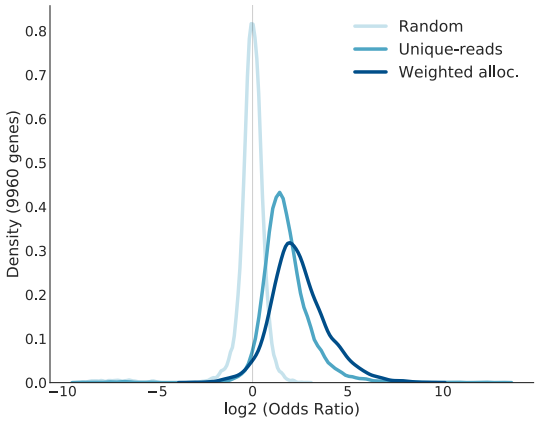
Weighted allocation		Maternal allele	
		Not expressed	Expressed
Paternal allele	Not expressed	1	6
	Expressed	0	279

(A)

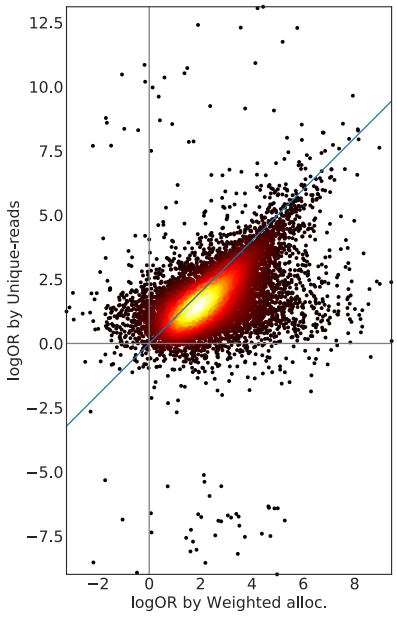
Unique-reads		Maternal allele	
		Not expressed	Expressed
Paternal allele	Not expressed	56	39
	Expressed	49	142



(B)



(C)



(D)

Figure 4

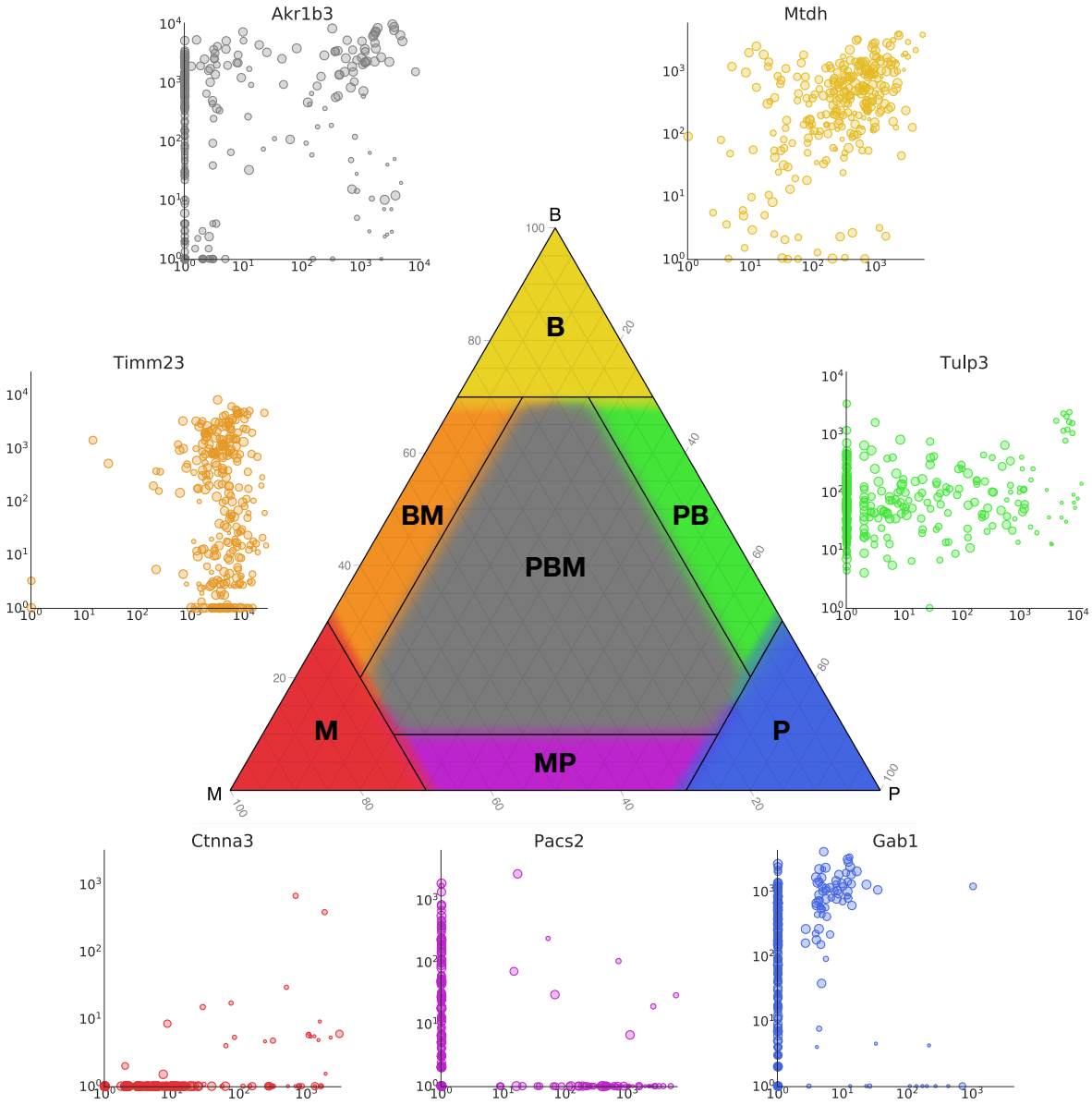


Figure 5

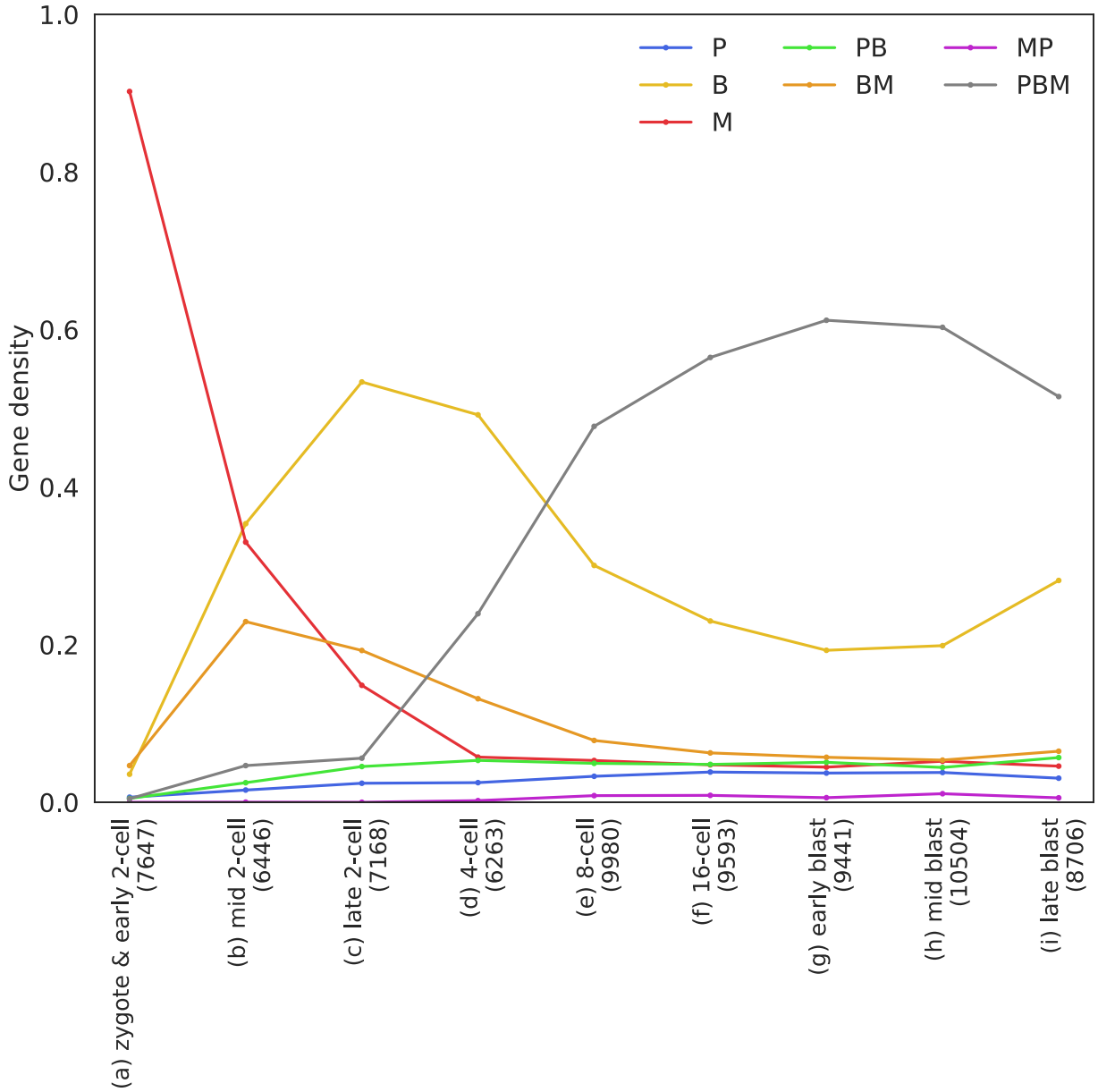
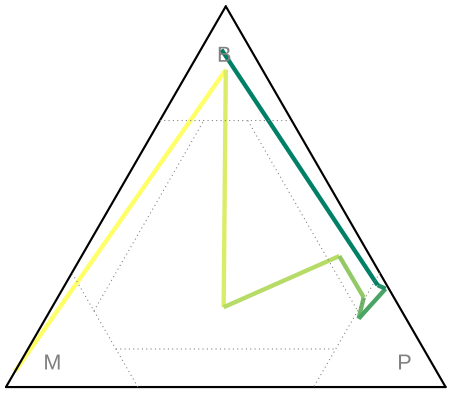
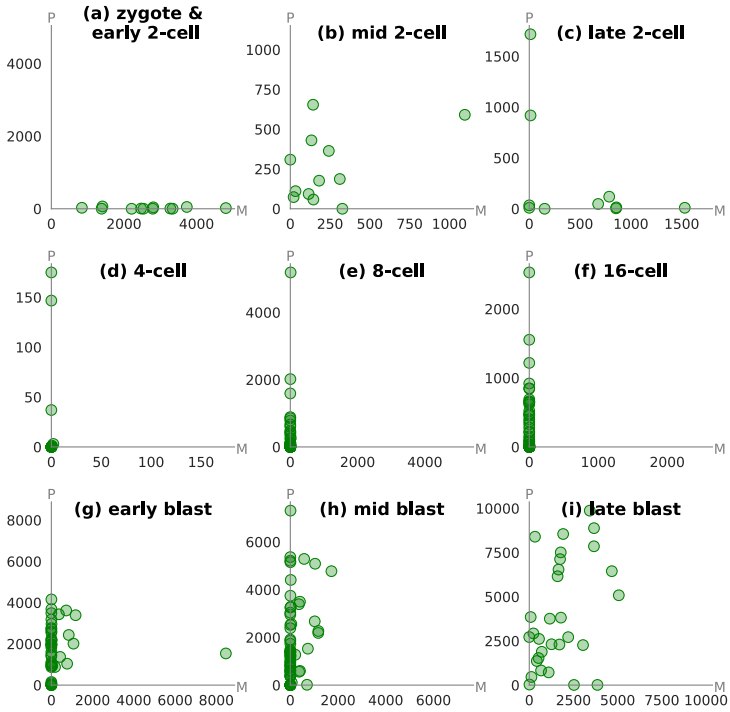


Figure 6



(A)



(B)

Figure 7

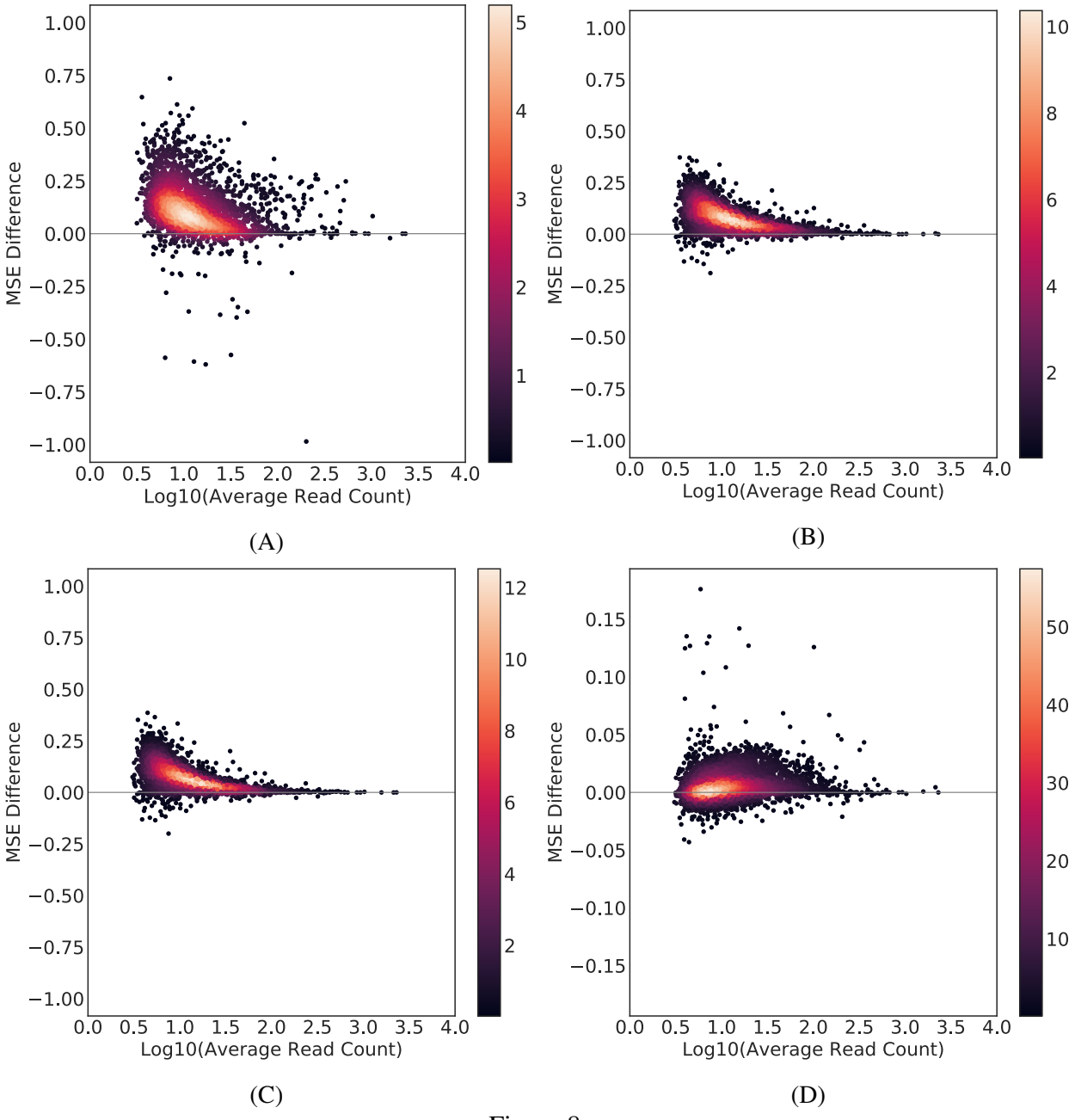


Figure 8

Supplemental Methods

Data

Deng [Deng et al., 2014] sampled 286 preimplantation embryo cells from an F1 hybrid of CAST×B6 along the stages of prenatal development. Embryos were manually dissociated into single cells using Invitrogen TrypLE and single-end RNA-Seq sequencing was performed using Illumina HiSeq 2000 (Platform GPL12112). We downloaded the data, Series GSE45719 from Gene Expression Omnibus (GEO) at http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE45719. There were fastq-format read files for 4 single-cell samples from zygote stage, 8 from early 2-cell, 12 from mid 2-cell, 10 from late 2-cell, 14 from 4-cell, 47 from 8-cell, 30 from 16-cell, 43 from early blastocyst, 60 from mid blastocyst, and 58 from late blastocyst stage.

Simulation

We randomly sampled 1% of reads in each of 60 cells at the mid blastocyst stage to obtain an average read count of ∼138k reads per cell. We repeated the sampling six times. We applied the weighted allocation algorithm to the full set of ∼13M reads and also applied each of four estimation methods to the down-sampled data. We compared estimates obtained from the down-sampled data to the full data estimates and computed the mean squared error of estimation for each gene.

scRNA-Seq read alignment

We reconstructed the CAST genome by incorporating known SNPs and short indels (Sanger REL-1505) into the reference mouse genome sequence (Genome Reference Consortium Mouse Reference 38) using `g2gtools` (http://churchill-lab.github.io/g2gtools/). We lifted the reference gene annotation (Ensembl Release 78) over to the CAST genome coordinates, and derived a CAST-specific transcriptome. The B6 transcriptome is based on the mouse reference genome. We constructed a bowtie (v1.0.0) index to represent the diploid transcriptome with two alleles of each transcript. We aligned reads using bowtie with parameters '–all', '–best', and '–strata', allowing for 3 mismatches ('-v 3'). These settings enable us to find all of the best alignments for each read. For example, if there is a zero-mismatch alignment for a read, then only and all alignments with zero mismatch will be accepted.

Weighted allocation of multi-reads

We applied the EMASE software [Raghupathy et al., 2018] to the diploid alignment profile of each individual cell to quantify read counts for maternal and paternal alleles of each gene. Isoform level counts were summed to obtain allele and gene-specific counts. EMASE is available at https://github.com/churchill-lab/emase.

Hierarchical mixture model: MCMC algorithm

We consider a single gene $\mathcal{G}$ in cells indexed by $k = 1, \ldots, K$. The maternal and paternal read counts, x_{gk} and y_{gk}, are estimated by the weighted allocation algorithm or unique-reads method. The expected counts are non-integral in many cases, we convert them to integer with floor function. The total read count for $\mathcal{G}$ is $n_{gk} \equiv x_{gk} + y_{gk}$. We define the maternal allelic proportion p_{gk} for each cell to be the expected proportion of maternal reads. We assume each cell is in one of three states with respect to the expression of $\mathcal{G}$: $s \in \{P, B, M\}$ where the indices denote (P)aternal monoallelic, (B)i-allelic, or (M)aternal monoallelic expression respectively. For each cell k, we introduce an indicator vector, $\boldsymbol{z}_{gk} = \left(z^P_{gk}, z^B_{gk}, z^M_{gk}\right)$ where

$$z^s_{gk} = \begin{cases} 1, & \text{if } \mathcal{G} \text{ is in expression state } s, \\ 0, & \text{otherwise} \end{cases} \qquad \text{and } \sum_s z^s_{gk} = 1. \tag{1}$$

By definition, the marginal distribution of $P(z^s_{gk} = 1) = \pi^s_{g\bullet}$ and $\sum_s \pi^s_{g\bullet} = 1$ and we interpret the mixture weights, $\pi^s_{g\bullet}$, to be the proportion of cells in allelic expression state s in the population. We propose a hierarchical mixture model:

$$\begin{aligned}
x_{gk} &\sim \begin{cases} \text{Binomial}(n_{gk}, p^P_g) & \text{if } z^P_{gk} = 1, \\ \text{Binomial}(n_{gk}, p^B_{gk}) & \text{if } z^B_{gk} = 1, \\ \text{Binomial}(n_{gk}, p^M_g) & \text{if } z^M_{gk} = 1 \end{cases} \\
p^P_g &\sim \text{Beta}(1, \alpha^{mono}_g) \\
p^B_{gk} &\sim \text{Beta}(\alpha^B_g, \beta^B_g) \\
p^M_g &\sim \text{Beta}(\alpha^{mono}_g, 1) \\
\alpha^{mono}_g &\sim \text{half-Cauchy}(7, 2) \\
\alpha^B_g,\, \beta^B_g &\sim \text{half-Cauchy}(2, 2) \\
\pi^s_{g\bullet} &\sim \text{Dirichlet}\left(\frac{1}{3}, \frac{1}{3}, \frac{1}{3}\right)
\end{aligned} \tag{2}$$

As analytical solution to the maximum likelihood estimators does not exist for these models, we implemented Markov-Chain Monte Carlo algorithm using `STAN` [Carpenter et al., 2017] version 2.17.0. While fitting `scBASE` model, we calculate π^s_{gk}

$$\begin{aligned}
\pi^P_{gk} = E(z^P_{gk}{=}1) &= \frac{\pi^P_{g\bullet}\,\text{Binomial}(x_{gk}|n_{gk}, p^P_g)}{N} \\
\pi^B_{gk} = E(z^B_{gk}{=}1) &= \frac{\pi^B_{g\bullet}\,\text{Binomial}(x_{gk}|n_{gk}, p^B_{gk})}{N} \text{ and} \\
\pi^M_{gk} = E(z^M_{gk}{=}1) &= \frac{\pi^M_{g\bullet}\,\text{Binomial}(x_{gk}|n_{gk}, p^M_g)}{N} \text{ where}
\end{aligned} \tag{3}$$

$$N = \pi^P_{g\bullet}\,\text{Binomial}(x_{gk}|n_{gk}, p^P_g) + \pi^B_{g\bullet}\,\text{Binomial}(x_{gk}|n_{gk}, p^B_{gk}) + \pi^M_{g\bullet}\,\text{Binomial}(x_{gk}|n_{gk}, p^M_g)$$

We then calculated p_{gk} in the following way.

$$p_{gk} = \pi^P_{gk}\,p^P_g + \pi^B_{gk}\,p^B_{gk} + \pi^M_{gk}\,p^M_g \tag{4}$$

Once model fitting finishes, we report a maximum a posteriori estimate of maternal allele proportion, $\tilde{p}_{gk}$, as the mean of sampled p_{gk}.

Hierarchical mixture model: Expectation-Maximization algorithm

For point estimation of maternal allele proportion in each cell, p_{gk}, for gene $\mathcal{G}$, we also propose an empirical Bayes model based on the same latent variable, $\boldsymbol{z}_{gk} = (z^P_{gk}, z^B_{gk}, z^M_{gk})$:

$$\begin{aligned}
x_{gk}\big|p_{gk}, n_{gk} &\sim \text{Binomial}(n_{gk}, p_{gk}) \\
p_{gk}\big|z^s_{gk} &\sim \text{Beta}(\alpha^s_g, \beta^s_g) \\
\boldsymbol{z}_{gk} &\sim \text{Multinomial}\Big(1, \{\pi^P_{g\bullet}, \pi^B_{g\bullet}, \pi^M_{g\bullet}\}\Big)
\end{aligned} \tag{5}$$

The distribution of $x_{gk}\big|n_{gk}, z^s_{gk}$ is Beta-Binomial and the full marginal distribution of $x_{gk}\big|n_{gk}$ is 3-component Beta-Binomial mixture:

$$\begin{aligned}
P(x_{gk}\big|n_{gk}) &= \sum_s P(z^s_{gk} = 1)\, P(x_{gk}|n_{gk}, z^s_{gk}{=}1) \\
&= \sum_s \pi^s_{g\bullet}\,\text{Beta-Binomial}\Big(x_{gk}\big|n_{gk}, \alpha^s_g, \beta^s_g\Big).
\end{aligned} \tag{6}$$

The log likelihood is

$$\begin{aligned}
\log L(\boldsymbol{X}_g|\boldsymbol{N}_g) &= \log P(x_{g1},\ldots,x_{gK}\big|n_{g1},\ldots,n_{gK}) \\
&= \log \prod_k P(x_{gk}\big|n_{gk}) \\
&= \sum_k \log \sum_s \pi^s_{g\bullet}\, P(x_{gk}|n_{gk}, z^s_{gk}=1) \\
&= \sum_k \log \sum_s \pi^s_{g\bullet} \binom{n_{gk}}{x_{gk}} \frac{B\big(\alpha^s_g + x_{gk},\, \beta^s_g + n_{gk} - x_{gk}\big)}{B\big(\alpha^s_g, \beta^s_g\big)} \\
\text{where } B(\alpha,\beta) &= \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}.
\end{aligned} \tag{7}$$

As analytical solution to the maximum likelihood estimators does not exist, we use the following Expectation-Maximization (EM) algorithm.

E-step:
Given parameters α^s_g, β^s_g, $\pi^s_{g\bullet}$, and data x_{gk} and n_{gk}, we can compute, π^s_{gk}, the posterior expectation on z^s_{gk}.

$$\begin{aligned}
\pi^s_{gk} &= E(z^s_{gk}=1|x_{gk}, n_{gk}, \widehat{\boldsymbol{\alpha}}_g, \widehat{\boldsymbol{\beta}}_g, \widehat{\boldsymbol{\pi}}_g) \\
&= \frac{P(z^s_{gk}=1)P(x_{gk}|n_{gk}, z^s_{gk}=1)}{\sum_{t\in\{P,B,M\}} P(z^t_{gk}=1)P(x_{gk}|n_{gk}, z^t_{gk}=1)} \\
&= \frac{\widehat{\pi}^s_{g\bullet}\, P(x_{gk}|n_{gk}, z^s_{gk}=1)}{\sum_t \widehat{\pi}^t_{g\bullet}\, P(x_{gk}|n_{gk}, z^t_{gk}=1)}
\end{aligned} \tag{8}$$

$$\text{where } \widehat{\boldsymbol{\alpha}}_g = \left\{\widehat{\alpha}^P_g, \widehat{\alpha}^B_g, \widehat{\alpha}^M_g\right\}, \widehat{\boldsymbol{\beta}}_g = \left\{\widehat{\beta}^P_g, \widehat{\beta}^B_g, \widehat{\beta}^M_g\right\}, \widehat{\boldsymbol{\pi}}_g = \left\{\widehat{\pi}^P_{g\bullet}, \widehat{\pi}^B_{g\bullet}, \widehat{\pi}^M_{g\bullet}\right\}, \text{ and,}$$

$$P\big(x_{gk}\big|n_{gk}, z^s_{gk}=1\big) = \binom{n_{gk}}{x_{gk}} \frac{B\big(\alpha^s_g + x_{gk},\, \beta^s_g + n_{gk} - x_{gk}\big)}{B\big(\alpha^s_g, \beta^s_g\big)}.$$

M-step:

Given the expectation on expression state of gene $\mathcal{G}$ in each cell, π^s_{gk}, we can compute the maximum likelihood estimators, $\widehat{\alpha}^s_g$, $\widehat{\beta}^s_g$, and $\widehat{\pi}^s_g$. What makes the maximum likelihood process more complicated in the proposed model is that n_{gk}'s are different cell by cell, and we want to put more weights on cells that have larger n_{gk}'s. Since π^s_{gk} describes how likely cell k belong to expression state s, we can make cells that are more likely to be in state s to contribute more in estimating parameters α^s_g, β^s_g, and π^s_g. We used the method of moments approach proposed in [Kleinman, 1973] in which he equates the weighted mean and variance of observed maternal allele proportion, x_{gk}/n_{gk}, to the analytical mean and variance of Beta-Binomial distribution re-parameterized with $\mu^s_g = \alpha^s_g/(\alpha^s_g + \beta^s_g)$ and $\tau^s_g = (1+\alpha^s_g+\beta^s_g)^{-1}$:

$$\begin{aligned}
E\left(\frac{x_{gk}}{n_{gk}}\middle| z^s_{gk}=1\right) &= \mu^s_g \\
Var\left(\frac{x_{gk}}{n_{gk}}\middle| z^s_{gk}=1\right) &= \frac{\mu^s_g(1-\mu^s_g)}{n_{gk}} + \tau^s_g\mu^s_g(1-\mu^s_g)\left(1-\frac{1}{n_{gk}}\right)
\end{aligned} \tag{9}$$

Then, parameters are updated in the following way.

$$\begin{aligned}\left(\widehat{\mu}_g^s\right)^{new} &= \bar{p}_g^s \\ \left(\widehat{\tau}_g^s\right)^{new} &= \frac{S_g^s - \bar{p}_g^s(1-\bar{p}_g^s)\left\{\sum_k \frac{w_{gk}^s}{n_{gk}}\left(1-\frac{w_{gk}^s}{W_g^s}\right)\right\}}{\bar{p}_g^s(1-\bar{p}_g^s)\left\{\sum_k w_{gk}^s\left(1-\frac{w_{gk}^s}{W_g^s}\right) - \sum_k \frac{w_{gk}^s}{n_{gk}}\left(1-\frac{w_{gk}^s}{W_{gk}^s}\right)\right\}}\end{aligned} \tag{10}$$

where

$$\begin{aligned}\bar{p}_g^s &= \frac{\sum_k w_{gk}^s\left(\frac{x_{gk}}{n_{gk}}\right)}{W_g^s} \\ S_g^s &= \sum_k w_{gk}^s\left(\left(\frac{x_{gk}}{n_{gk}}\right) - \bar{p}_g^s\right)^2, \\ w_{gk}^s &= \pi_{gk}^s \cdot \frac{n_{gk}}{1+(n_{gk}-1)\left(\widehat{\tau}_g^s\right)^{old}}, \text{ and} \\ W_g^s &= \sum_k w_{gk}^s\end{aligned} \tag{11}$$

From updated μ_g^s and τ_g^s, we can easily derive the parameters that are used in (5).

$$\begin{aligned}\left(\widehat{\alpha}_g^s\right)^{new} &= \left\{\frac{1}{\left(\widehat{\tau}_g^s\right)^{new}} - 1\right\}\left(\widehat{\mu}_g^s\right)^{new}, \text{ and} \\ \left(\widehat{\beta}_g^s\right)^{new} &= \left(\frac{1}{\left(\widehat{\tau}_g^s\right)^{new}} - 1\right)\left(1-\left(\widehat{\mu}_g^s\right)^{new}\right)\end{aligned} \tag{12}$$

Here we designed the weight of each cell, w_{gk}^s, to reflect both π_{gk}^s and $1/\tau_g^s$. By definition, π_{gk}^s gets larger when the gene $\mathcal{G}$ in k-th cell is more likely to be in expression state s. Similarly, $1/\tau_g^s$ increases if the observed allele proportion, x_{gk}/n_{gk}, reside in narrower range across the cell population, which positively correlates to the sample sizes, n_{gk}. Therefore, cells that are more likely to be in state s and also having higher depth of coverage would contribute more on estimating parameter α_g^s and β_g^s. Finally, we update $\pi_{g\bullet}^s$ by adding the expected expression state of $\mathcal{G}$ across the cell population.

$$\left(\widehat{\pi}_{g\bullet}^s\right)^{new} = \frac{1}{K}\sum_k \pi_{gk}^s \tag{13}$$

After parameter values converge, we also report the posterior expected allele proportion, $\tilde{p}_{gk}(\equiv \tilde{x}_{gk}/n_{gk})$. We show here that the posterior allele proportion of cell k lies in-between what is observed in cell k itself and what is found from other cells in the same expression state s.

$$\begin{aligned}\tilde{p}_{gk} &= E\left(x_{gk}\middle|n_{gk}\right) \\ &= E\left(\sum_s \pi_{g\bullet}^s \text{Beta-Binomial}\left(x_{gk}\middle|n_{gk},\alpha_g^s,\beta_g^s\right)\right) \\ &= \sum_s \pi_{g\bullet}^s \frac{\widehat{\alpha}_g^s}{\widehat{\alpha}_g^s + \widehat{\beta}_g^s}\end{aligned} \tag{14}$$

Model implementation in scBASE *package*

The single-component Beta-Binomial model

```
data {
    int<lower=1> N;
    int<lower=0> n[N];
    int<lower=0> x[N];
}
parameters {
    real<lower=0> kappa;
    real<lower=0,upper=1> phi;
    vector<lower=0,upper=1>[N] theta;
}
model {
    kappa ~ cauchy(0, 2);
    theta ~ beta(phi*kappa, (1-phi)*kappa);
    x ~ binomial(n, theta);
}
```

The three-component Beta-Binomial mixture model

```
data {
    int<lower=1> N;
    int<lower=0> n[N];
    int<lower=0> x[N];
}
parameters {
    simplex[3] pi;
    real<lower=7> a_mono;
    vector<lower=2>[2] alpha;
    vector<lower=0,upper=1>[N] theta;
}
model {
    vector[3] log_pi = log(pi);
    a_mono ~ cauchy(7, 2);
    alpha ~ cauchy(2, 2);
    theta ~ beta(alpha[1], alpha[2]);
    for (i in 1:N) {
        vector[3] lps = log_pi;
        lps[1] = lps[1] + beta_binomial_lpmf(x[i]|n[i], a_mono, 1);
        lps[2] = lps[2] + beta_binomial_lpmf(x[i]|n[i], 1, a_mono);
        lps[3] = lps[3] + binomial_lpmf(x[i]|n[i], theta[i]);
        target += log_sum_exp(lps);
    }
}
generated quantities {
    matrix[N,3] pi_z;
    real log_sum_exp_log_pi_z_raw;
    for (i in 1:N) {
        vector[3] log_pi_z_raw = log(pi);
        log_pi_z_raw[1] = log_pi_z_raw[1] + beta_binomial_lpmf(x[i]|n[i], a_mono, 1);
        log_pi_z_raw[2] = log_pi_z_raw[2] + beta_binomial_lpmf(x[i]|n[i], 1, a_mono);
        log_pi_z_raw[3] = log_pi_z_raw[3] + binomial_lpmf(x[i]|n[i], theta[i]);
        log_sum_exp_log_pi_z_raw = log_sum_exp(log_pi_z_raw);
        for (j in 1:3)
            pi_z[i,j] = exp(log_pi_z_raw[j] - log_sum_exp_log_pi_z_raw);
    }
```

```
}
```

Classification of a gene according to its ASE profile across many cells

We classify a gene according to the proportion of cells in P-, B-, and M-states, $\left(\pi_{g\bullet}^{P}, \pi_{g\bullet}^{B}, \pi_{g\bullet}^{M}\right)$, that are estimated by the partial pooling model. If a majority of cells ($\pi_s > 0.7$) are in a particular ASE state, $s \in \{P, B, M\}$, then we will assign the gene to the class **P** (monoallelic paternal; blue), **B** (biallelic; yellow), or **M** (monoallelic maternal; red) respectively. Otherwise, when a majority of cells ($\sum \pi_s > 0.9$) are a mixture of two of those classes, we classify it into either of **PB** (mixture of monoallelic paternal and biallelic; green), **BM** (mixture of monoallelic maternal and biallelic; orange), or **MP** (a mixture of monoallelic maternal and paternal; purple). There were genes that all three ASE states appear, and we put them in **PBM** (mixture of all; gray). We specified these seven classes in a ternary simplex diagram (Figure 5). The class boundaries are somewhat arbitrary and those are only a suggestion.

References

[Carpenter et al., 2017] Carpenter, B., Gelman, A., Hoffman, M., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M., Guo, J., Li, P., and Riddell, A., *et al.*, 2017. Stan: A probabilistic programming language. *Journal of Statistical Software, Articles*, **76**(1):1–32.

[Deng et al., 2014] 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**(6167):193–196.

[Kleinman, 1973] Kleinman, J. C., 1973. Proportions with extraneous variance: Single and independent sample. *Journal of the American Statistical Association*, **68**(341):46–54.

[Raghupathy et al., 2018] Raghupathy, N., Choi, K., Vincent, M. J., Beane, G. L., Sheppard, K. S., Munger, S. C., Korstanje, R., Pardo-Manual de Villena, F., and Churchill, G. A., 2018. Hierarchical analysis of rna-seq reads improves the accuracy of allele-specific expression. *Bioinformatics*, **34**(13):2177–2184.

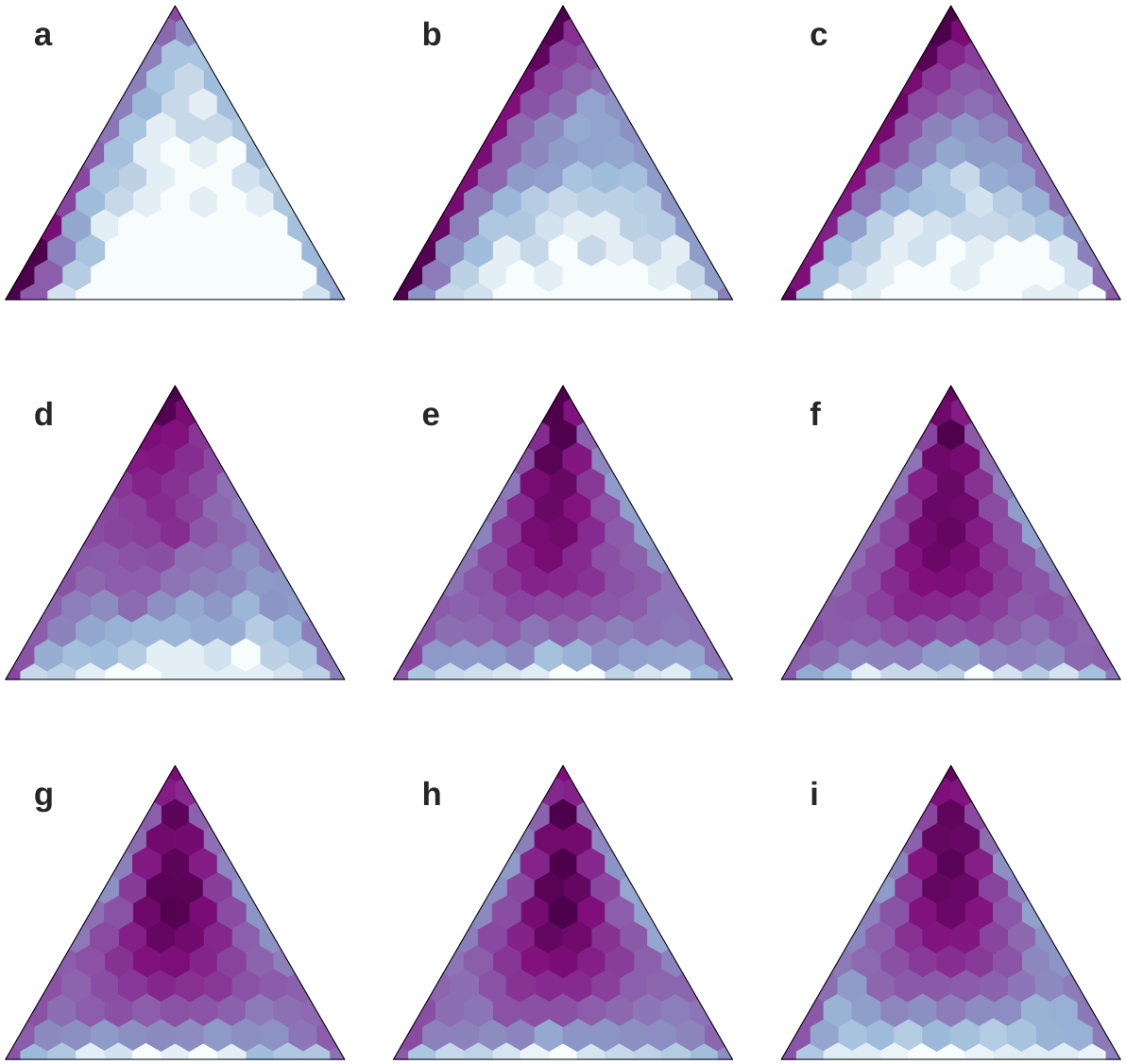
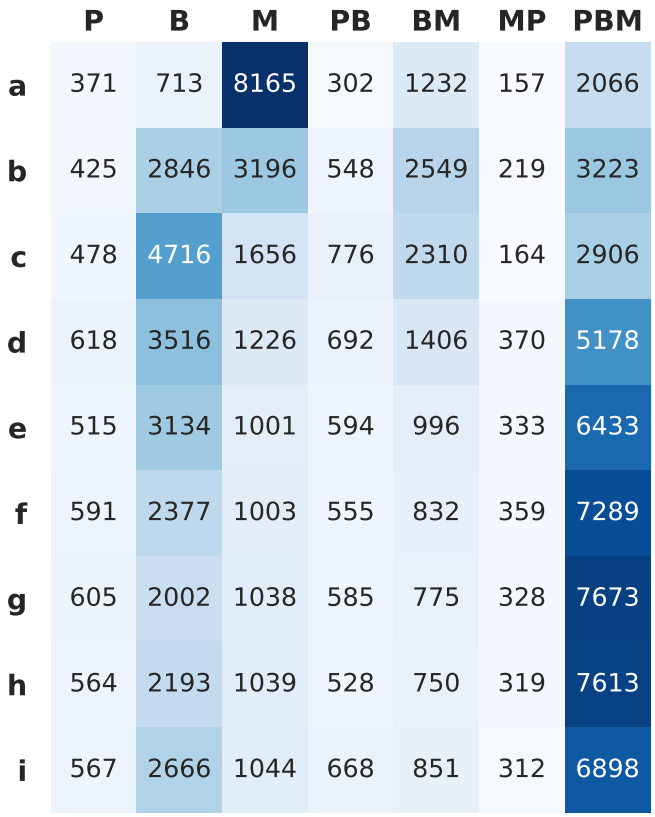


Figure S1: Heatmaps show the distribution of genes over the simplex diagram for **a** the zygote & early 2-cell, **b** mid 2-cell, **c** late 2-cell, **d** 4-cell, **e** 8-cell, **f** 16-cell, **g** early blastocyst, **h** mid blastocyst, and **i** late blastocyst stage of development



	P	B	M	PB	BM	MP	PBM
a	371	713	8165	302	1232	157	2066
b	425	2846	3196	548	2549	219	3223
c	478	4716	1656	776	2310	164	2906
d	618	3516	1226	692	1406	370	5178
e	515	3134	1001	594	996	333	6433
f	591	2377	1003	555	832	359	7289
g	605	2002	1038	585	775	328	7673
h	564	2193	1039	528	750	319	7613
i	567	2666	1044	668	851	312	6898

Figure S2: The number of genes in each class along the developmental stages (See Figure 6).

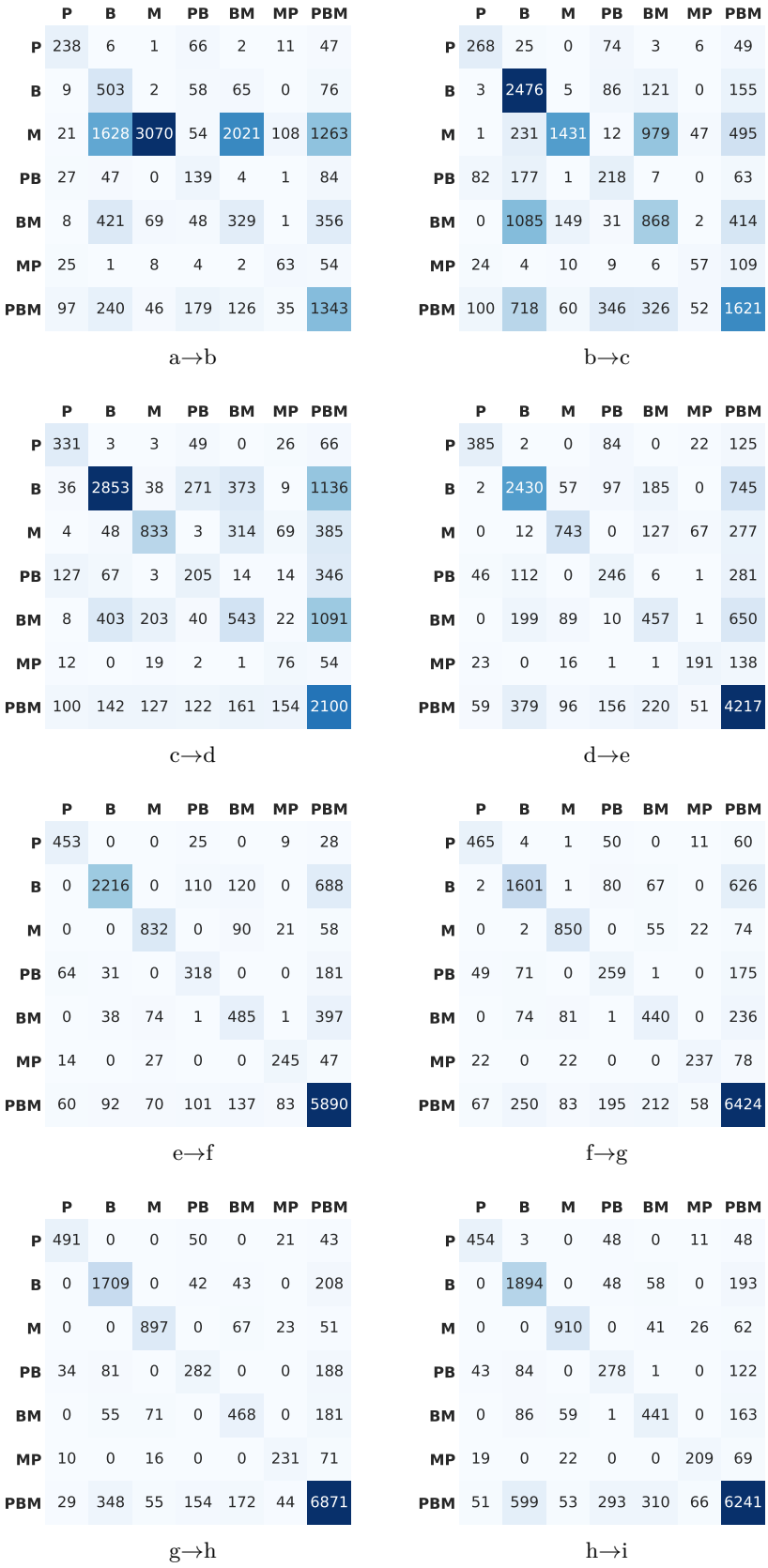


Figure S3: The number of genes that change their class along the developmental stages (See Figure 6 and Supplemental Figure S2).