Allele Specific Expression from Single Cell RNA-Seq Data

The Jackson Laboratory
Leading the search
for tomorrow's cures

Kwangbom Choi, Narayanan Raghupathy, Steve C. Munger, Gary A. Churchill*
The Jackson Laboratory, Bar Harbor, Maine 04609 USA

BACKGROUND

Single-cell RNA-Seq (or scRNA-Seq) data poses multiple technological challenges that result from amplification of the cellular RNA and limited size of the RNA pool in individual cells. High level of sampling noise often coupled with low depth of coverage causes larger variance in quantified allele expressions which often causes overdispersion issues for our analysis models. Due to inefficient, semi-random reverse transcription process, alleles may *drop out* from measurements irrespective of their abundance. In addition, low expressed alleles are subject to *sampling zero* from underlying Poisson process. What makes it challenging is that technical dropout or sampling zero of one allele and random monoallelic expression (RME) of the other allele would look exactly same in our observations. Sorting out the combined effect of technical dropouts and sampling noise on top of cell-to-cell stochasticity at the allele level is key to understanding the genetic regulation and dynamics of allele transcription from scRNA-Seq data.

Genes	C57BL/6J	CAST/EiJ
Cdk2ap1	**0**	**408**

(a)

Genes	C57BL/6J	CAST/EiJ
Cdk2ap1	31569	31398
Gm12184	29889	29889
Trim7	2415	2415
Gm4252	4	4
Gm29485	1145	1145

31699

(b)

Genes	C57BL/6J	CAST/EiJ
Cdk2ap1	2187	
Gm12184	99	
Trim7		
Gm4252		
Gm29485	1	

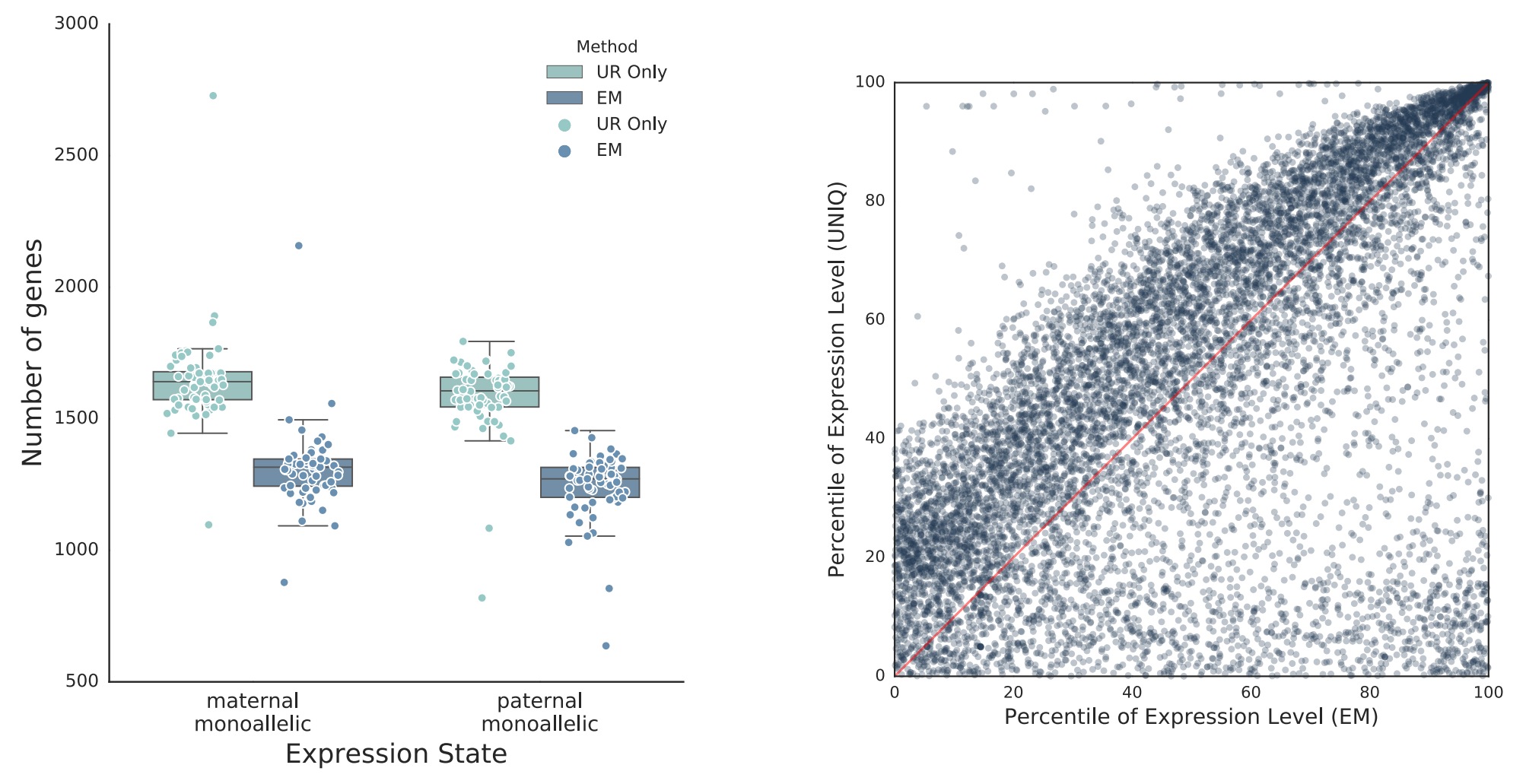
(c)

Genes	C57BL/6J	CAST/EiJ
Cdk2ap1	579	
Gm12184		
Trim7		
Gm4252		
Gm29485		

(d)

Genes	C57BL/6J	CAST/EiJ
Cdk2ap1	**16421.8**	**14350.3**
Gm12184	651.8	651.8
Trim7	0	0
Gm4252	0	0
Gm29485	0.6	0.6

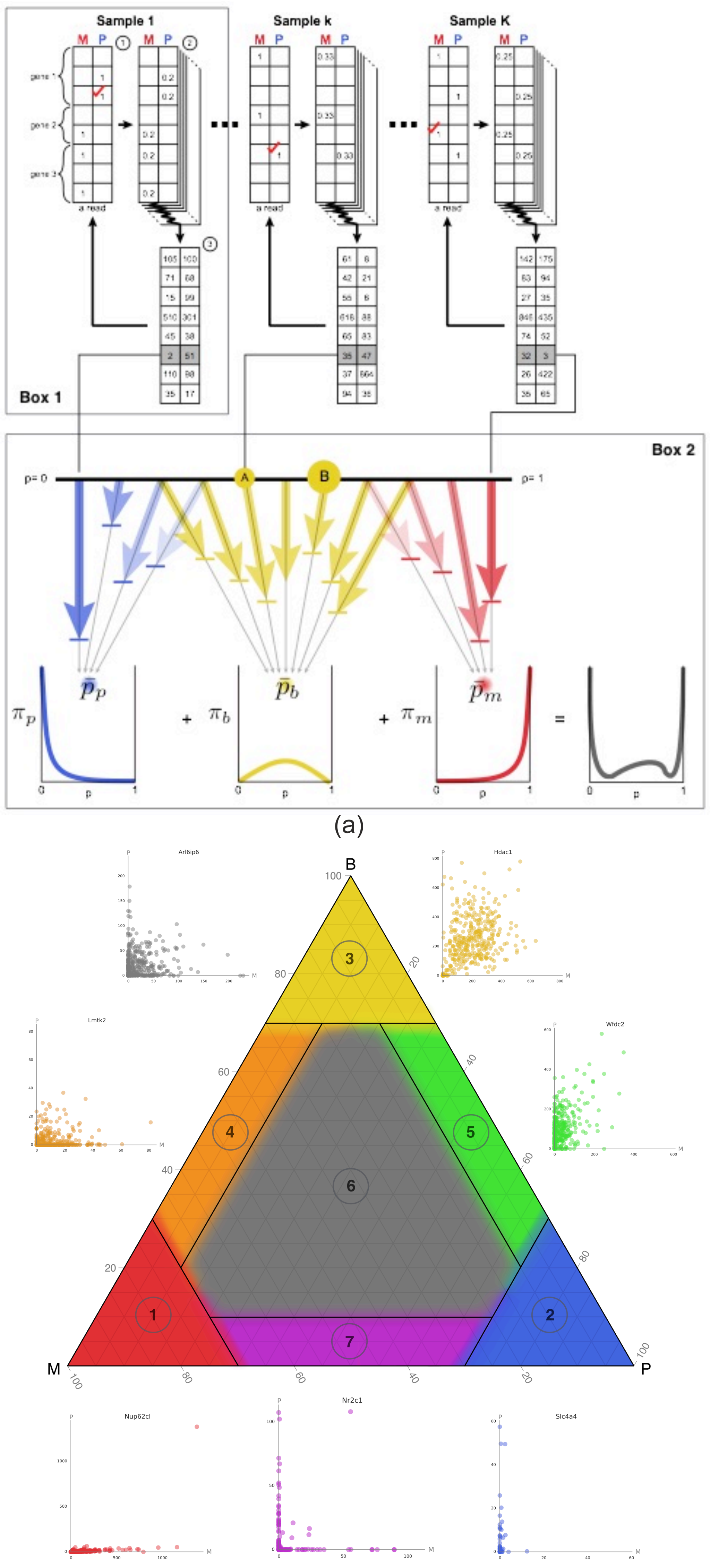
(e)



(f) (g)

Figure 1: (a) According to the number of unique reads, Cdk2ap1 is seemingly expressed only from CAST/EiJ (or CAST) allele. (b)~(d) In fact, there were many reads that aligned to only to a specific group. Especially, there were 579 reads align uniquely to the C57BL/6J (or B6) alleles of Cdk2ap1 and Gm12184. All multireads exclusively aligned to a specific group, and they are ambiguous but still informative for ASE quantification. (e) Final EM result. Our alignment data strongly supports that B6 allele of Cdk2ap1 is expressed: Cdk2ap1 is not CAST monoallelic gene. (f) The number of genes identified to be monoallelic expression between methods using UR-only (light blue) and all the multireads (dark blue). Over 300 genes were false monoallelic expression in UR-only strategy. (g) Percentile difference between expression based on UR-only and EM using all the multireads. Some genes have many more multireads than others, and therefore, discarding multireads can influence the rankings on expression levels. There existed over 400 genes dramatic (over 50%) drop in percentile if we discard reads just because they aligned to multiple location of genome.

OUR APPROACH



(a)

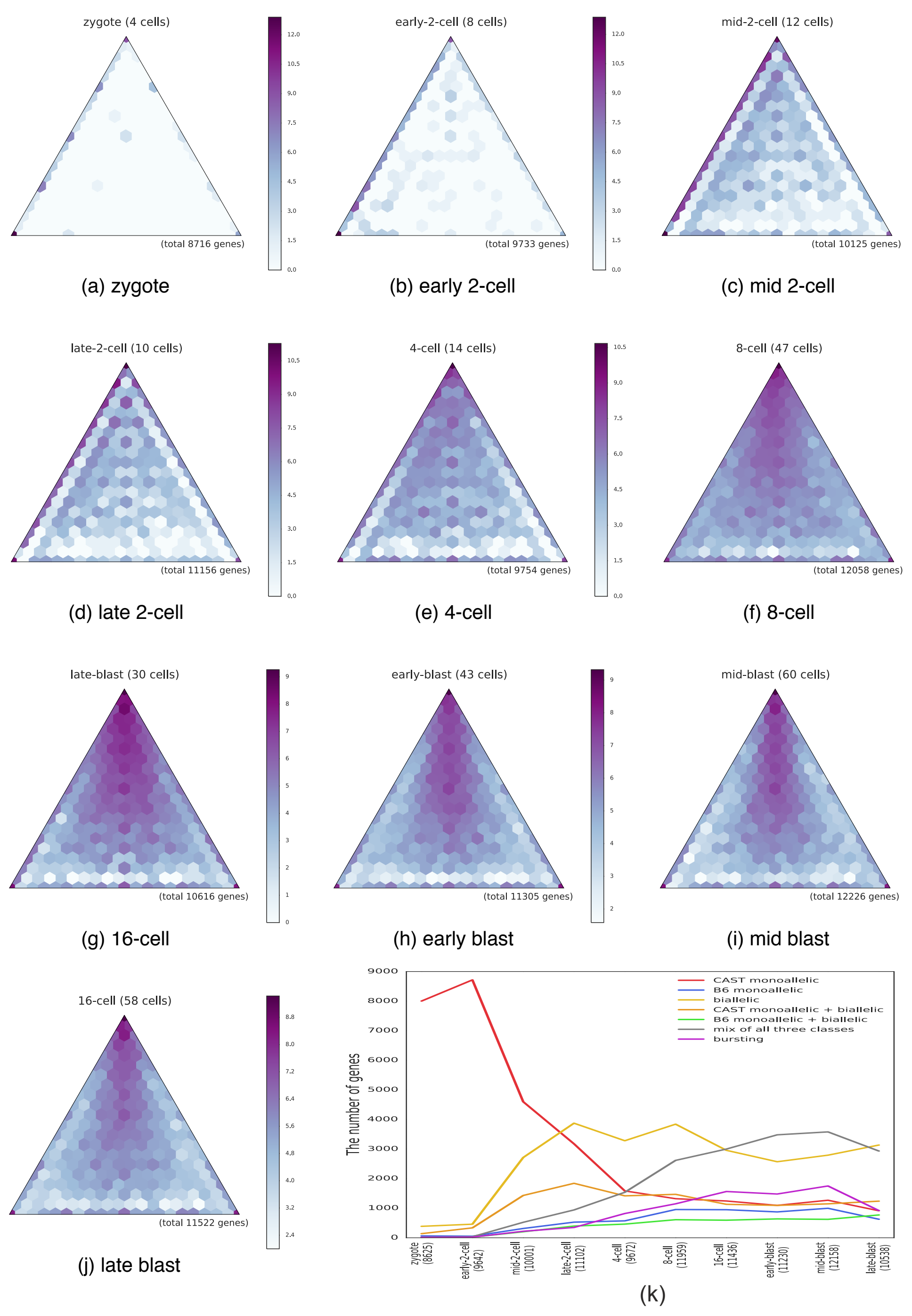
(b)

Figure 2: (a) An overview of our two-step approach, scASE. **Step 1:** For each transcript, quantify expected allele-specific read counts. If a read aligned to multiple alleles of transcripts (①), we disambiguate its origin by deriving posterior expectation that represents how likely a read originate from a particular allele of a transcript it aligned to (②). The expected allelic read counts are obtained by summarizing posterior expectations across all reads (③). As we can compute better expected read counts when we have better expectation of read origin and vice versa, we alternate ② and ③ until both converge. **Step 2:** Adjust maternal allele proportion (p_M) derived from allele-specific read counts we estimated in Step 1 by combining information from other cells in similar expression state. (b) A schematic diagram of gene classification on the simplex of Maternal monoallelic (π_m), Paternal monoallellic (π_p), and Biallelic cell proportions. Our scASE model classifies genes according to cell proportion $[\pi_m , \pi_p , \pi_b]$ of those three ASE patterns. Class 1 and 2 are genes that most (over 70%) of cells express either preferentially from maternal (red) or paternal allele (blue). Class 3, 4, 5, and 6 are genes that presumably transcribe from both alleles but non-negligible (over 10%) proportion of cells behave as if they are monoallelic. We speculate that Class 4 and 5 are genes that have allelic imbalance since, for example, we would see more maternal allele missed from dropouts or sampling zeros on genes that preferentially express paternal allele. Allelic expression in Class 7 genes are mutually exclusive: most of cells (over 90%) are either maternal or paternal monoallelic.

DATA

We tested our model using a published dataset of Deng et al [1]. They sampled total of 286 preimplantation embryo cells from F1 hybrid of CAST/EiJ×C57BL/6J along the stages of prenatal development: from zygote to early 2-cell, mid 2-cell, late 2-cell, 4-cell, 8-cell, 16-cell, early, mid and late blastocyst. 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 entire dataset, Series GSE45719, from Gene Expression Omnibus (GEO).

RESULTS



(a) zygote (b) early 2-cell (c) mid 2-cell
(d) late 2-cell (e) 4-cell (f) 8-cell
(g) 16-cell (h) early blast (i) mid blast
(j) late blast (k)

Figure 3: The ASE pattern of cell population changes over stages of blastocyst development. Initial predominance of maternal allele gradually weakens while more genes are preferentially or simultaneously expressing paternal alleles. In later stages, (f)-(j), most genes transcribe from both alleles although there still exist cells that transcribe only from either maternal or paternal alleles. The number of genes in each class of population ASE pattern is shown in (k).

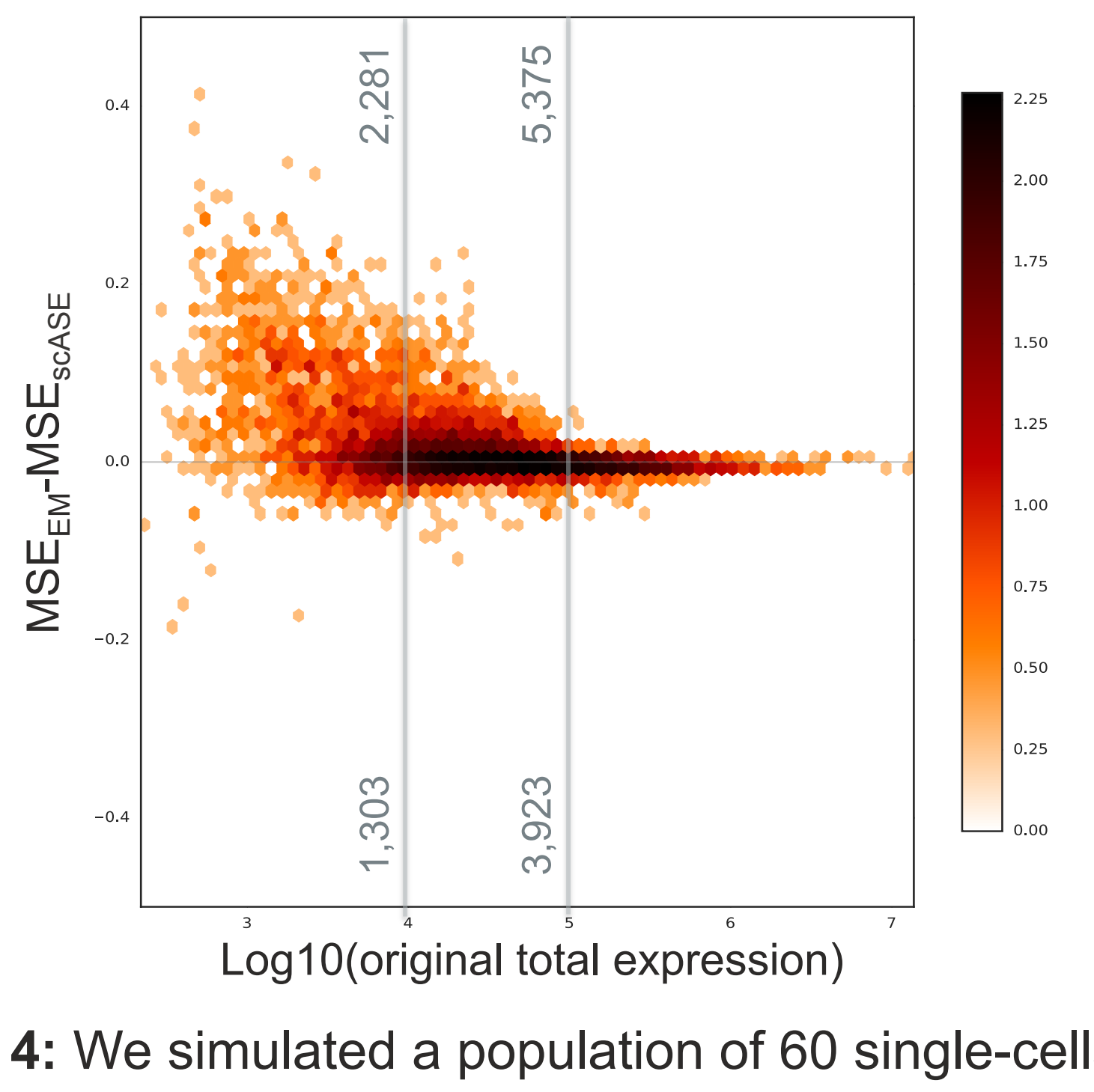


Figure 4: We simulated a population of 60 single-cells by sampling 10% of reads (avg 1.3M reads per cell) from the original data (mid-blast stage) and tested how close our model estimates are to the EM expected read counts of full read set. Overall, our model is more stable for estimating ASE from genes of low expression by partially combining information across cell population.

REFERENCE

[1] Q. Deng et al. (2014) Single-cell reveals dynamic, random monoallelic gene expression in mammalian cells. Science. **343**:193-196.

www.jax.org