Bioinformatics, 36, 2020, i251–i257
doi: 10.1093/bioinformatics/btaa474
ISMB 2020

BIRD: identifying cell doublets via biallelic expression from single cells

Kerem Wainer-Katsir and Michal Linial*

Department of Biological Chemistry, The Institute of Life Sciences, The Hebrew University of Jerusalem, Jerusalem, Givat Ram 91904, Israel

*To whom correspondence should be addressed.

Abstract

Summary: Current technologies for single-cell transcriptomics allow thousands of cells to be analyzed in a single experiment. The increased scale of these methods raises the risk of cell doublets contamination. Available tools and algorithms for identifying doublets and estimating their occurrence in single-cell experimental data focus on doublets of different species, cell types or individuals. In this study, we analyze transcriptomic data from single cells having an identical genetic background. We claim that the ratio of monoallelic to biallelic expression provides a discriminating power toward doublets' identification. We present a pipeline called Blallelic Ratio for Doublets (BIRD) that relies on heterologous genetic variations, from single-cell RNA sequencing. For each dataset, doublets were artificially created from the actual data and used to train a predictive model. BIRD was applied on Smart-seq data from 163 primary fibroblast single cells. The model achieved 100% accuracy in annotating the randomly simulated doublets. Bonafide doublets were verified based on a biallelic expression signal amongst X-chromosome of female fibroblasts. Data from 10X Genomics microfluidics of human peripheral blood cells achieved in average 83% ($\pm$3.7%) accuracy, and an area under the curve of 0.88 ($\pm$0.04) for a collection of $\sim$13 300 single cells. BIRD addresses instances of doublets, which were formed from cell mixtures of identical genetic background and cell identity. Maximal performance is achieved for high-coverage data from Smart-seq. Success in identifying doublets is data specific which varies according to the experimental methodology, genomic diversity between haplotypes, sequence coverage and depth.

Contact: michall@cc.huji.ac.il

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Single-cell RNA sequencing (scRNA-seq) technology has evolved very rapidly in recent years (Kolodziejczyk *et al.*, 2015; Lan *et al.*, 2017; Zheng *et al.*, 2017; Zilionis *et al.*, 2017). scRNA-seq enables higher resolution of the expression profile of specific cells within cells tissue and enables accurate assessment of the single cells' identity and variability. This new technology has been applied for a wide range of biological studies and across many organisms, including rodents and humans. Complex tissues were dissociated and sequenced by scRNA-seq resulting in cataloging cells by their types determining tissue composition (Usoskin *et al.*, 2015; Villani *et al.*, 2017; Zeisel *et al.*, 2015) and identifying overlooked cell types (Buettner *et al.*, 2015).

All scRNA-seq studies rely on profiling cell transcriptomes. The main hurdle in obtaining reliable and high-quality data from scRNA-seq stems from the limited amounts of RNA per cell and the stochastic nature of transcription (Ilicic *et al.*, 2016). Specifically, most current scRNA-seq methods suffer from low capture efficiency and high dropouts (Haque *et al.*, 2017). Additionally, all single-cell expression data are signified by a strong signal of monoallelic expression, which is not detected from the sequencing pool of cells. The dominant monoallelic expression of single cells (Borel *et al.*, 2015; Jiang *et al.*, 2017) is attributed to allelic dropout of transcripts due to the insufficient coverage, and to the cellular phenomenon of 'transcriptional burst' (Reinius and Sandberg, 2015). The latter means that each of the alleles has its kinetics, thus at any specific time expression mostly stems from a single allele (Larsson *et al.*, 2019).

Innovative technologies for scRNA-seq were developed to increase high throughput while minimizing biological intrinsic and technical errors [discussed in Bacher and Kendziorski (2016), Chen *et al.* (2019) and Hashimshony *et al.* (2016)]. Some methods make use of fluorescence-activated cell sorting (Kolodziejczyk *et al.*, 2015; Wagner *et al.*, 2018) and microfluidic-based platforms, such as the C1 Single-Cell Auto Prep System (Fluidigm) (Xin *et al.*, 2016). These methods are usually followed by a full-length transcript sequencing as in Smart-seq2 (Picelli *et al.*, 2014). Others use droplet microfluidic procedures that combine a tagging step before cell lysis [reviewed in Chen *et al.* (2019) and Klein *et al.* (2015)]. Advances in the droplet technique allow capturing beads with a single cell per droplet (dscRNA-seq) thus increasing the scale for single-cell

transcriptomic by two orders of magnitude (Fan *et al.*, 2015; Sheng *et al.*, 2017). These protocols use only poly-A sequencing and are thus biased toward the 3′ side of the transcript. Most current-day protocols include additional steps of barcoding the transcripts by unique molecular tag identifiers (Klein *et al.*, 2015), and further improvement of the capturing efficiency (Sheng and Zong, 2019).

One of the pitfalls in the field concerns a faulty identification of a doublet of cells as a single cell. Doublets rate depends on the concentration of the input cells, which is estimated from the dilution Poisson statistics (Macaulay *et al.*, 2017). An increase in doublets rate is also associated with the unique features of the subjected tissue and cells' isolation protocols. New methods for increasing cell capturing that reduces costs include multiplexing protocols (Zheng *et al.*, 2017). By increasing the number of cells as input, the multiplexed droplet RNA-Seq (dscRNA-seq) benefits from reducing technical noise (Zhang *et al.*, 2019). However, as a byproduct, it leads to an unavoidable increase in the number of cell doublets. One of the methods to identify the rate of doublets in the data includes mixing cells from different origin [e.g. rodents and human (Zheng *et al.*, 2017)]. Alternatively, the dscRNA-seq setting was carried over single cells from several individuals with a different genomic background that were intentionally mixed for estimating the fraction of doubles in the sample (Kang *et al.*, 2018). Benefiting from the SNP profile of each an individual, the Demuxlet algorithm was applied to estimate mixed individual doublets (Kang *et al.*, 2018). A recently published Scrublet algorithm analyzes single cells for identifying problematic multiplets according to the nearest neighbor graph-based classifier (Wolock *et al.*, 2019). DoubletFinder (McGinnis *et al.*, 2019a) benefits from the unique cell-state expression profiles and identify doublets from transcriptionally distinct cells. While these set of methods can differentiate cell mixtures from distinct individuals and cell types, they do not attempt to differentiate cells that originate from the same source or cell type.

In this study, we analyze data from scRNA-seq and dscRNA-seq for identifying doublets without any prior knowledge of cell-type composition. Instead, we take advantage of monitoring allele-specific expression biases. The method called BIallelic Ratio for Doubles (BIRD) relies on analyzing heterologous SNPs present in scRNA-seq data. We report on the accuracy on identifying doublets, which is strongly dependent on sequencing methodologies, coverage, depth and the degree of allelic diversity in the genomic data.

2 Materials and methods

2.1 Dataset of single cells

2.1.1 Dataset 1: primary human fibroblasts

A dataset of scRNA-seq of female fibroblast UCF1014 was downloaded from the European Genome-phenome Archive (https://www.ebi.ac.uk/ega/home) using accession number EGAD00001001083. The data consist of two sets of scRNA-seq: 104 cells (22 PCR cycles) and 59 cells (12 PCR cycles). The data were collected in a C1 Auto Prep System (Fluidigm) device (Xin *et al.*, 2016) and sequenced using full transcript Smart-seq2 (Picelli *et al.*, 2014). DNA-seq of UCF1014 was also downloaded from EGAD00001001084. The sequence data were produced and described by Borel *et al.* (2015).

2.1.2 Dataset 2: peripheral human blood mononuclear cells

The data were created and described in Kang *et al.* (2018). Peripheral blood mononuclear cells (PBMCs) scRNA-seq from eight different individuals were downloaded from the Gene Expression Omnibus database, accession number GSE96583. This dataset contains three different runs. Two of the runs include a mixture of scRNA-seq from four different individuals (run_a and run_b sets). The third run is a mixture of all eight individuals scRNA-seq data (run_c). Cells were sequenced using 10X Genomics (Chromium instrument) methodology. Additional VCF files of exome sequencing of these individuals were extracted through Github link (https://github.com/yelabucsf/demuxlet_paper_code/tree/master/fig2). It shares also an additional file determining the individuals' origin per each scRNA-seq as processed by the Demuxlet tool (Kang *et al.*, 2018). Only cells that were assigned by Demuxlet to belong to the same individual and therefore could not be explicitly annotated as singlets or doublets were used for further analysis by our methodology.

2.2 Biallelic score for single cells

To correctly estimate the allelic-specific expression (ASE) and specifically, the degree of biallelic expression of each cell, the DNA-seq of cell line UCF1014 was used to create a collection of all heterozygous SNPs (hSNPs) using Gene Analysis Toolkit (Van der Auwera *et al.*, 2013). All the hSNPs were kept in a VCF file. The RNA-Seq reads were preprocessed using Trimmnomatic (Bolger *et al.*, 2014) with its default parameters. Using STAR (Dobin and Gingeras, 2015) each scRNA-seq FastQ file was aligned against the GRCh37 (hg19) UCSC female reference (Ref) (after excluding Y-chromosome). For dataset 1 (see Section 2.1) The BAM output of the alignment and the hSNPs VCF were processed using Allelcounter-master (Castel *et al.*, 2015). The tool creates a table containing the number of reads for each SNP that matches the Ref and the Alternative (Alt) alleles. Then, we process the table into two tables, one for the Ref alleles and the other for the Alt alleles. Both tables contain the number of reads assigned to each cell for each hSNP. An observation is considered for cells having $\geq$6 reads for a specific hSNP in a specific cell. The same procedure was applied for all single-cell datasets analyzed (described in Section 2.1).

For the PBMC dscRNA-seq data (Dataset 2) BAM files of the three runs were split to single cell BAM files to maintain an individual-based BAM file per cell. Each single cell was identified according to its unique cell-based barcode. Each BAM file was coupled to its corresponding individual VCF according to the identification by the Demuxlet algorithm. (Kang *et al.*, 2018) and was preprocessed by Allelcounter-master (Castel *et al.*, 2015). We then unified all cells that share a specific run, for a specific individual into two tables containing the number of reads for each hSNP that matches the Ref and the Alt alleles. As each individual contains its own set of hSNPs, tables for the Ref, Alt were created for each of the 16 run-individual pair. For this analysis, for a hSNP to be considered, we require the number of reads to be $\geq$3 per hSNP for the subjected cell.

For both datasets we calculate for every available hSNP the allelic ratio (AR) of that hSNP in a specific cell as:

$$\mathrm{AR}_{hc} = \left\{ \frac{\text{Alt reads}}{\text{Alt reads} + \text{Ref reads}} \right\}_{hc},$$

where h refers to hSNP and c to a specific cell. The AR ranges between 0 and 1, with a minimal value of 0.0001 for all Ref allele. For a hSNP with no evidence for expression, the value is zero. Value of 1 is associated with all hSNPs that are fully aligned to the Alt allele. Genuine biallelic hSNP are bounded by the AR values ($0.1 \leq \mathrm{AR} < 0.9$).

An allele independent score for biallelic ratio (BAR) was calculated as follows: For a given cell and a given gene, let i be an index of the informative (heterozygous) variants, and define by Ref_i and Alt_i the number of Ref and Alt reads each informative variant. Define by $\mathrm{Tot}_i = \mathrm{Ref}_i + \mathrm{Alt}_i$ the total number of reads for the variant, and by $\mathrm{Min}_i = \min\{\mathrm{Ref}_i, \mathrm{Alt}_i\}$ the minimal number of reads out of the two alleles of the variant. Let $i_* = \mathrm{argmax}_i(\mathrm{Min}_i)$ be the most informative variant with the maximal BAR (for the given cell and gene combination). We then define the BAR of the cell-gene as:

$$\mathrm{BAR} = \frac{\mathrm{Min}_{i*}}{\mathrm{Tot}_{i_*}}.$$

Then, for each cell we take the average BAR of all its expressed genes. In a formal notation

$$\text{Biallelic Ratio}_{cg} = \frac{\mathrm{Max}\{\mathrm{Min}\{\text{Ref reads}_{ic}, \text{Alt reads}_{ic}\}\}_{gc}}{\{\text{Ref reads}_{igc} + \text{Alt reads}_{igc}\}}$$

i-hSNP location (in the numerator stands for the specific SNP that was Max in the denominator), c stands for cell and g for a gene.

2.3 Doublet simulation and validation

To create a Ref dataset of doublets, we created doublets *in silico* for each of the analyzed datasets separately. For the simulations we randomly sample 10% of the single cells to be mixed into cell doubles. The other 90% of single cells remain singles. This process eventually creates a composed collection with 5% of the original cells being simulated doublets. The pair mixing is done by summing together the cells' reads from the Ref and Alt tables. Following summation, for the fibroblast data (Dataset 1), we randomly down-sample the reads to the average cell reads number. Due to the low coverage of the PMBCs data (Dataset 2) we skipped this step. In each simulation, we record the BAR values for the singlets and the simulated doublets. The procedure of creating simulated doublets was repeated 100 times. For each run, we also record the average of the BAR values for all the singlets and the average of all simulated doubles.

The primary fibroblasts of Dataset 1 originated from female (Borel *et al.*, 2015) and further processed as in Wainer-Katsir and Linial (2019). Thus, we used the unique property of X-inactivation to obtain an expression pattern that matches the cell specific activation of one of the X-chromosomes (Garieri *et al.*, 2018). Specifically, we calculate AR per hSNP per cell. Then, we calculate AR* for assessing the BAR for Chromosome X. We consider AR* to be 1-AR in cases that AR>0.5, thus $0 \leq AR^* \leq 0.5$.

$$AR* = \begin{cases} if\ AR \leq 0.5\ AR \\ if\ AR > 0.5\ 1 - AR \end{cases}$$

AR* balances between the hSNP expression from either the Ref or Alt alleles. To avoid a noisy signal from a sporadic expression of hSNP, we consider only SNPs that were transcribed in >25% of the cells. We also removed hSNPs that were fully monoallelic to the Ref or the Alt allele (i.e. AR*<0.1). Out of these hSNPs, for each cell, we calculate the average of AR*. Cells with an average AR* score of >0.05 show an unexpected biallelic X-chromosome expression and are thus considered suspicious as doublets.

2.4 Statistical measures for cell doublet identification

For both datasets Mann–Whitney U test was used to determine differences between singlets and simulated doubles according to the BAR values. For Dataset 1 that is based on Smart-seq2, we applied a Gaussian Mixture Model (GMM) that differentiates the groups of singlets from doublets. The GMM was set with two components one seeking the singlets and the other the doubles. The features that were given to the GMM include (i) the BAR of each of the cells, and (ii) the number of expressed genes in heterozygous sites in each of the cells.

Dataset 2 (based on 10X Genomics technology) is signified by a poor coverage; therefore, we included additional features per cell for recovering doublets. The four features that were used are: (i) the amount of reads over all heterozygous positions; (ii) the number of expressed genes having heterozygous positions; (iii) the average BAR values; and (iv) the fraction of genes defined as biallelic out of all genes expressed in that cell. Each of these features was standardized according to the specific run-individual pair (Kang *et al.*, 2018). Each dataset was standardized and trained accordingly on its own values. The datasets were split to training and test sets (with the training set covers 75% of the data). We applied random forest (RF) procedure with the four listed features for recording the statistical results. Operating the RF classifier was done with the following parameters: n_estimators =100, random_state =42, min_samples_leaf =sqrt(sample size), min_samples_split =2*sqrt(sample size). Additionally, singlets and doublets were equal weighted by demanding the class_weight to be balanced. Sensitivity, specificity and accuracy in doublets identification were measured according to the success and failure in detecting simulated and candidate doubles. Receiver operating characteristics (ROC) curve and area under the curve (AUC) were calculated for each run-individual couple, each of the different runs (run_a, run_b and run_c) and for the combined set of all three runs.

2.5 Cell separation by gene expression matrix

In this study, we followed the protocol in Lun *et al.* (2016b). Count matrix of genes over cells was created for each of the samples using HTSeq (Anders *et al.*, 2015). The genes to cells matrix was analyzed using SingleCellExperiment Package (Risso *et al.*, 2018), scater package (McCarthy *et al.*, 2017) and scran (Lun *et al.*, 2016). Rtsne package was used to create the t-distributed stochastic neighbor embedding (t-SNE) (Pezzotti *et al.*, 2017) representation of the 26 first principal components of the PCA of the gene expression profile of each individual and each run.

3 Results

3.1 Overview BIRD pipeline

In single-cells transcriptomics, monoallelic expression of alleles across each of the heterozygous positions is a common phenomenon (Fig. 1A). The majority of the hSNPs are monoallelic due to the stochastic nature of expression (Borel *et al.*, 2015; Reinius and Sandberg, 2015). In doublets, if one cell expresses one of the alleles and the others cell the other, the result is shift toward the biallelic expression (i.e. $0.1 \leq AR < 0.9$). Therefore, a signal with AR centered around 0.5 represents a product of expressing hSNPs derived from both alleles. The key concept underlying BIRD is that doublets can be identified by a signal derived from the shift toward higher BAR (see Section 2). The transformation of each gene and each cell from AR to its average BAR value is illustrated in Figure 1B, left. The distribution of BAR values from all cells is indicative for the presence of cells that display a substantial biallelic expression and thus are most likely cell doublets (Fig. 1B, right). Testing the performance of BIRD to identify doublets, is based on artificially creating doublets by combining expression profiles from random single cells and testing the potency of statistical methods to correctly identify such *in silico* simulated doublets (Fig. 1C and D).

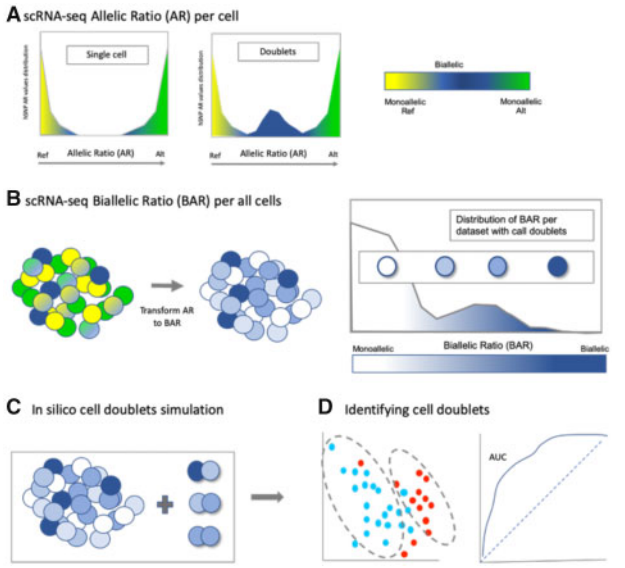

Fig. 1. (left) Illustration of the BIRDs scheme for scRNA-seq and dscRNA-seq data. (**A**) Illustrative schemes for the distribution of AR calculated per each cell. AR values range between 0 and 1, for the Ref (yellow) and Alt (green) alleles. The blue corresponds to biallelic expression. (i) For single cells, AR reflecting an apparent monoallelic expression; (ii) cell collection with doublets is signified by a shift in AR to around 0.5. (**B**) For every gene in every cell, the AR is estimated. BAR for every gene is calculated. BAR values for scRNA-seq that includes doublets is shown. The BAR is bounded between 0 and 0.5 from monoallelic (white) to biallelic expression (dark blue). (**C**) Simulation of randomly selected single-cells pairs is performed to create a dataset composed from the original and simulated cell doubles. (**D**) ML statistical technique differentiating singlets from doublets. The success of doublet identification is assessed by visualization and standard measures (e.g. AUC)

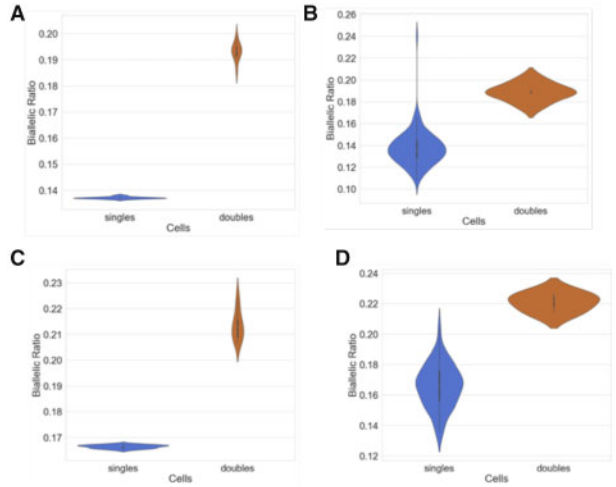

Fig. 2. (Right) BAR values for human single cells primary fibroblasts population and *in silico* simulated doublets. Doublets simulation was done for two datasets of human primary fibroblasts, which differ by the number of PCR cycles used prior to sequencing. (**A, B**) 104 cells, 22 PCR cycles and (**C, D**), 59 cells, 12 PCR cycles. Violin plots of the BAR mean values for all single cells means versus simulated doubles means based on 10 simulations (A, C) and cell means of a single simulation (B, D). The statistical test used is Mann–Whitney U test. The results are (A) statistic =0, *P*-value $<e-34$; (B) *P*-value =$4.25e-30$; (C) statistic =0, *P*-value =$1.281e-34$; and (D) *P*-value =$1.43e-17$

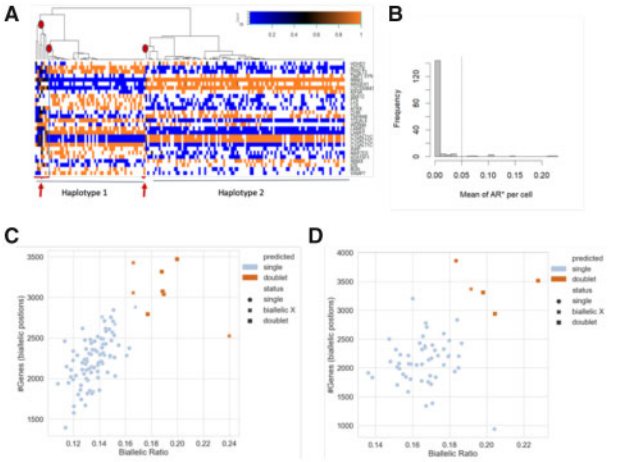

Fig. 3. Validation of cell doublets according to X-chromosome allelic specific pattern. Dataset is a combined collection of human fibroblasts (163 cells). (**A**) A matrix of AR values of cells (columns) and hSNPs (rows) is shown (without simulation). Only high-support hSNPs are included (see Section 2). Gene names are listed according to their chromosomal order (right). The hSNPs are colored from blue (AR=0, Ref allele) to orange (AR=1, Alt allele), with darker colors marking biallelic expression. A hierarchical clustering of the cells indicates two main haplotypic origin (Haplotypes 1 and 2). The arrows and the circles above the branches of the clustering tree indicate cells with strong biallelic expression across the X-chromosome. (**B**) Histogram of cells as in (A), by their AR* values. AR*=0 indicates monoallelic X-chromosome expression, and larger value marks a higher biallelic expression level. The dashed line is a natural threshold separating X-inactivated monoallelic from biallelic expressed hSNPs. (**C, D**) Scatter plots show the success of detecting singlets from doubles following a single simulated set for the 104 cells (C) and 59 cells (D). Symbols correspond to singlets as circles and doublets as squares. A cell is marked by x if it was validated as a doublet on the background of the X-inactivated status (as in A, B). Dark orange marks cells that were identified by the GMM classifier (see Section 2) as doubles

3.2 BAR values for the fibroblast scRNA-seq data

The human primary fibroblast cells (total of 163 single cells) are comprised of two datasets according to the PCR protocol used for creating the sequencing library. The first collection consists of 104 cells that underwent 22 PCR cycles, and the second set consists of 59 cells that underwent 12 PCR cycles. Due to the different PCR protocols, the sequencing depth is different between the two cell collections and they are thus treated for the BIRD protocol as independent sets. Figure 2 shows the results of doublet simulations for each of these datasets. The distributions of singlets versus doubles for 100 simulation runs (each with 5% of artificially created doublets) are shown in Figure 2A and C. Both datasets resulted in a perfect separation with (Mann–Whitney U test statistic =0, and *P*-value $<e-34$). The mean values for a single simulation run are also very significant (Fig. 1B and D). The results from the 104 cells (22 PCR cycles) and the 59 cells (12 PCR cycles) show Mann–Whitney U test with a *P*-value of $4.25e-30$ and $1.43e-17$, respectively. The results of identifying doublets are data-specific but highly significant for the two cell collections despite the different sequencing depth associated with each.

3.3 Doublets verification based on Chromosome X-inactivation expression pattern

The primary fibroblast cells are of a female origin. Thus, in each cell, only one of the two X-chromosomes is active (i.e. Xa) while the other is inactivated. The expression patterns for the subset of hSNPs with substantial evidence are shown in Figure 3A. Most cells (columns) are signified by a single expression pattern that is indicated by Haplotype 1 and Haplotype 2. Only a few cells lean toward biallelic expression pattern over many X-chromosome genes. Based on hierarchical clustering of the cells, the cells that are suspicions as doublets are clustered in the leftmost subtree and on the leftmost leaf of the other two subtrees. The distribution of the AR* values for all 163 cells is shown in Figure 3B. AR*=0 means monoallelic X-chromosome expression, and the higher the AR*, the higher the biallelic expression is. Applying a natural threshold that separates cells with monoallelic and biallelic patterns (the striped line at AR*=0.05) allows focusing on cells that cross the threshold (eight cells). These cells are marked as cell doublet candidates. Notably, these suspicious eight cells are also signified by a higher BAR values for the 104 (Mann–Whitney U test *P*-value =$2.94e-4$) and 59 (Mann–Whitney U test *P*-value =0.023) cells.

3.4 Unsupervised identification of doublets for the fibroblast scRNA-seq data

GMM was used to separate singles from doubles. For the means of the 100 simulations, the separation between the singles and the doubles means reached 100% accuracy. For an illustrative of a single simulation run the mean BAR (*x*-axis) of each cell is plotted with the number of genes that are expressed in biallelic positions (*y*-axis, Fig. 3C and D). The scatter-plots symbols represent cells that are singlets, candidate cell doublets according to Chromosome X biallelic expression, and artificial doublets that are created by *in silico* simulations for the two fibroblast cell collections (104 and 59 cells based on PCR protocol for 22 and 12 cycles, respectively). Cells that were predicted as doublets by the GMM (whether true or false) are shown (dark orange). It is evident that most doublets and the Chromosome X candidate doublets have relatively high BAR values and are classified as doublets. For the 104 single-cells dataset, all simulated doublets (total 5) were identified (average of 5, SD of 0, with 100% detection rate), and 3/6 (average of 2.28, SD of 0.54, with 38% detection rate) of the cell candidates by X biallelic expression were identified. For the 59 cells, all 3 simulated doublets were identified (average of 2.84, SD of 0.62, with 94.66% detection rate) and 50% of the cell candidate doublets (average of 1.32, SD of 0.69, with 66% detection rate) according to the Chromosome X biallelic expression were correctly identified.

3.5 BAR values for the PBMCs dscRNA-seq data

The 13 364 PBMCs originate from 16 datasets that account for a pair of a run and an individual. When compared to the fibroblast cell collections (Dataset 1, see Section 2), the dscRNA-seq is

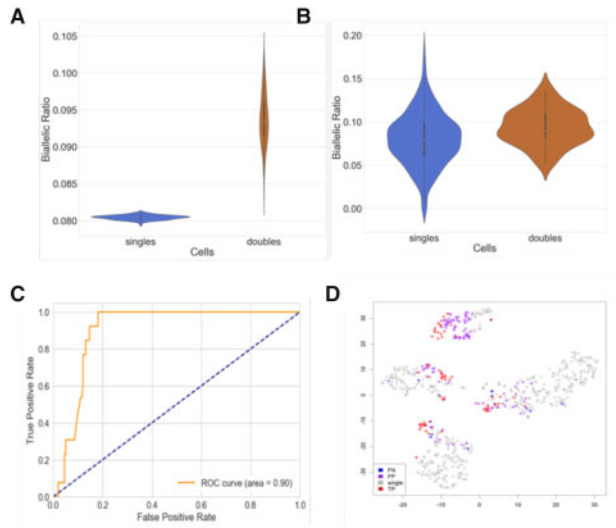

Fig. 4. Data used are dscRNA-seq marked as b_1493 based on Dataset 2. Violin plots of the mean of BAR values for single cells and the *in silico* simulated doubles. The mean BAR values shown were tested by Mann–Whitney U test for (**A**) 100 simulations (statistic = 0, *P*-value $<e-34$), and (**B**) a single simulation set (*P*-value = $1.22e-03$). (**C**) A ROC curve is shown based on a RF model fitting on a simulated dataset of singles and doubles. (**D**) t-SNE classification on PCA reduced data of cell expression. Each dot represents a single cell or a simulated doublet. Singlets (gray) correspond to cells that are singles and were predicted by the model as singles. TP (red) correspond to cells that are simulated as doublets and correctly predicted as such. False negatives (blue) are simulated doubles that were missed by the model. FP (purple) are misclassified by the model as doubles

characterized by a much lower coverage. Specifically, the number of informative genes is >2000 for the fibroblasts and only about 25 on average for the PBMCs.

BIRD was run on each of the 16 run-individual pairs and then *in silico* simulations created 5% of the samples to be doublets. The results from activating the BIRD process for an individual representative (run_b, individual 1493, denoted b_1493, 766 cells) are shown in Figure 4.

Figure 4A shows the violin plots of the mean of the BAR values for 100 simulations for singlets when compared to cell doublets. The separation is maximal with Mann–Whitney U test yields a statistic = 0, and a *P*-value $<e-34$. For a single run of the simulation (Fig. 4B), the trend of the doublets being more biallelic is kept with a separation of the Mann–Whitney U test yielding a *P*-value of $1.22e-03$. Note that, using BAR values alone is insufficient to distinguish between singles and doubles due to the high intrinsic noise in the data originated by the 10x Genomics protocol.

3.6 Supervised identification of doublets in the PBMC dataset

Including additional features and applying a supervised RF machine learning (ML) protocol (see Section 2), we reached a perfect separation with 100% accuracy when mean values of singlets and mean values of simulated doubles are compared for 100 simulation runs. For a single simulation run, we show the results of the ROC curve for the unseen, disjoint test set (Fig. 4C), with an AUC of 0.88 (SD 0.04) based on 10 simulation runs. We exploit the expression profile for the cell collection of b_1493 sample to create a t-SNE representation (Fig. 4D). The cells are color coded according to the prediction results. Note that, most identified doublets are positioned at the border of the expression clusters, but eventually other predicted doublets are fully embedded within an expression cluster. Recall that the expression profile information was not used by BIRD protocol for the prediction.

Similar to the analysis performed for a single dataset (b_1493), we repeated the analysis for all 16 combinations of runs and individuals. The AUC for the different runs of run-individual pairs were

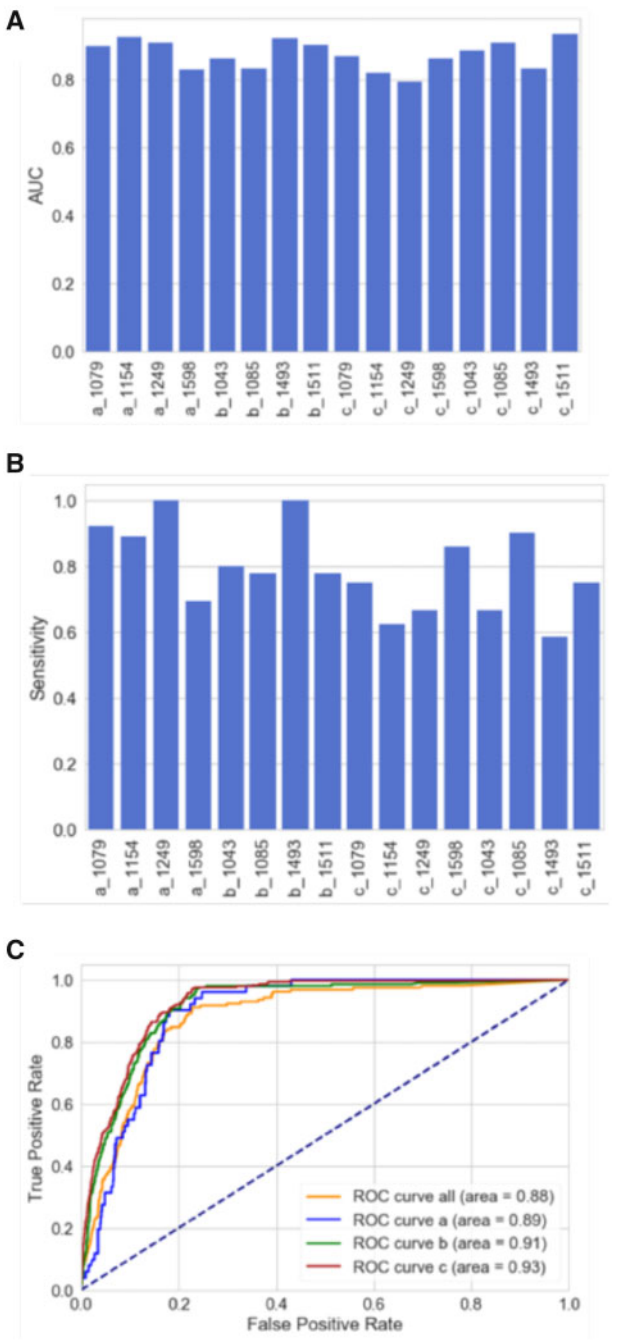

Fig. 5. Success in identifying cell doublets simulated from multiplex 10X experiment covering 13 364 single cells. (**A**) AUC for the test sets of each of run-individual couples. Runs refer to run_a, run_b that consist of four different individuals each and run_c that combines all eight individuals. (**B**) The sensitivity achieved for the test set for each of the tested individuals. (**C**) ROC curve is shown based on a RF model fitting on a simulated dataset of singles and doubles for all the cells (marked all) and for each of the separated runs (a, b and c)

combined to present the data for each of the runs and the entire dataset (with 13 364 cells, Fig. 5A). We tested whether the performance (as indicated by the AUC) is a mere reflection of the number of cells. However, it is evident that the success in identifying doublets and the number of cells that are associated with each dataset are not correlated. The sensitivity [i.e. TP/(TP+FN)] for each of the 16 datasets is shown in Figure 5B. The doubles rate in the sample (including the simulated cells) is shown in. Each sequencing runs is unique in term of coverage per individuals and the number of cells involved,

we created a unified ROC curve per each run and determined the AUC associated with each run and the whole dataset (Fig. 5C).

The t-SNE cell representations according to BIRD prediction based on the RF protocol for each of the 16 datasets in all instances, the t-SNE representation shows that accurate predictions [true positives (TP)] tend to cluster together with cells that are marked as false positives (FP). The estimate for the fraction of doublets from the mixture of two individuals is ∼5% (Kang *et al.*, 2018). Therefore, we expect many of the cells that are marked as FP to be doublets that are naturally present in the original data.

4 Discussion

Collecting data of single-cell transcriptomes had exposed a new dimension of cell variability. This technology had a direct impact on a wide range of biological questions across all domains of life (Stegle *et al.*, 2015). Some of these questions are sensitive to the faulty annotation of singlets as doubles or vice versa. While the presence of unrecognized cell doublets from the same cell type will not influence the misinterpretation for new cell types (Usoskin *et al.*, 2015; Villani *et al.*, 2017; Zeisel *et al.*, 2015), it might jeopardize interpretation concerning transcription regulation including transcriptional bursting kinetics (Larsson *et al.*, 2019), Chromosome X-inactivation phenomenon (Garieri *et al.*, 2018; Tukiainen *et al.*, 2017), escaping from it (Wainer-Katsir and Linial, 2019) and more.

We describe BIRD as a computational/statistical method that enables the identification of cell doublets from scRNA-seq data. The method complements other methods that rely on detailed cell mixing and cell-type expression profiles. BIRD method takes advantage of the BAM files generated for each scRNA-seq. The ASE that is extracted from the BAM files is often unused. This is since, routinely, the post-sequencing analysis starts with a cell to gene matrix representation thus discards the allelic information. BIRD takes advantage of this transparent feature for identifying doublets.

Recent methodologies present their potency toward the task of doublet identification. These methods are based on adding a pre-sequencing biochemical modification step for barcoding cells. Such non-trivial tagging procedures exploit antibodies to common cell surface antigens (cell hashing) (Stoeckius *et al.*, 2018). The antibody-based method is applicable to cells that carry the relevant antigens. A new method (MULTI-seq) successfully uses lipid modification step for cell indexing. It was shown to be eligible for solid tissues and frozen cells (McGinnis *et al.*, 2019). While there are many advantages for tagging cells before cell lysis and sequencing, an additional step in the experimental design can lead to batch effect and other technical and experimental biases. In contrast, the computational method is generic, yet data sensitive. BIRD shows no preference to the identity of the expressed genes, to the specific cell type or any of the cell extraction protocol.

We illustrate the high *performance* of BIRD mostly on *in silico* simulated doublets. Other studies estimated the occurrence of doublets by artificial mixing of cells of multiple types of cells from different organisms (Kang *et al.*, 2018; McGinnis *et al.*, 2019; Zheng *et al.*, 2017). However, when a solid tissue is treated to produce a collection of single cells, the protocol must overcome the adhesion forces between cells, extracellular matrix cohesion and more. Additionally, some cells tend to aggregate and clump following their isolation. All these technical issues may lead to an increasing number of doublets from neighboring cells with identical genetic background and expression profile. Therefore, current estimates for doublet contamination based on peripheral blood samples may be misleading. We anticipate that the number of reported doublets of cell mixtures from solid tissue is underestimated and can now be estimated using BIRD.

There are a number of limitations of BIRD protocol that need to be addressed: (i) the method is dependent on pre-knowledge of the individual genomics for assigning hSNPs from the sequenced scRNA-seq. With the fast accumulation of whole-genome and exome sequencing in humans and other model animals, it is anticipated not to be a limiting factor in the near future. (ii) The assessment of doublets using the notion of Chromosome X-inactivation is only valid for cells of female origin. Furthermore, for 50% of the cases, cells can be mixed without providing a biallelic signature (i.e. a mixture of the same Xa haplotype). (iii) While BIRD protocol ignores the gene expression profile, a scenario in which the profiles of cell mixtures do not overlap with each other can occur. This will result in cell doublets that do not contribute different alleles of the same genes and thus will not increase the BAR values. In such cases, BIRD protocol lacks the power to identify doublets. The use of other doublet cell identification is advisable (e.g. McGinnis *et al.*, 2019; Wolock *et al.*, 2019). (iv) The method relies on the dominant properties of stochasticity in the allelic expression of cells. Datasets that are far less stochastic might display a higher biallelic signal. Under such conditions, the ability to detect doublets is masked. The monoallelic fraction in single cells is a variable property of the experiment (Kim *et al.*, 2015) with mouse cells showing a lower degree of ASE relative to humans (Deng *et al.*, 2014; Tang *et al.*, 2011). Indeed, testing the scRNA-seq from F1 mice strains (Larsson *et al.*, 2019) confirm that the BAR value distribution is consistent with an intrinsic biallelic signature (not shown). In such cases, there is a need to employ BIRD only on genes that exhibit a more stochastic property and are signified by a monoallelic expression.

Overall, we described two types of datasets that use the BAR feature for discriminating singlets from doublets. The overall coverage of hSNPs and sequence depth are drastically different among the two analyzed datasets. The unsupervised GMM tool was sufficient in separating singles from doublets in a dataset of high-hSNP coverage (based on Smart-seq2 technology, with a full transcript sequencing). The high coverage of this methodology is fundamental to the ability of BIRD to robustly testing a biallelic signal despite a limited number of cells (163 cells, Dataset 1). The other dataset (dataset 2, 10X Genomics) yields shallow coverage, which is restricted to the 3′ tail of the transcripts and was therefore trained using RF. Despite the described coverage and difference in the sequencing protocols (3′ based versus a full-length transcript), the mean values of the simulated doublets were 100% identifiable for the larger 104 cells dataset and 94.66% for the smaller 59 cells set indicating higher BAR values for doublets with respect to singlets. The noisy and sparse data associated with 10X Genomics remain challenging and should be assessed for each run.

In summary, BIRD protocol is a generic method to identify doublets according to the deviation of single-cells biallelic profile. For most of the genomics 10X datasets that include hundreds of cells each, AUC on the task of identifying doublets for the three different runs reached an average value of 0.88 (±0.04). We applied BIRD on datasets of different coverage, scale and accuracy. BIRD uses a data-driven protocol and is applicable in all instances where in addition to the scRNA-seq/dscRNA-seq data genomic heterologous sites can be extracted.

Funding

This work was partially supported by #9960 from Yad Hanadiv.

Conflict of Interest: none declared.

References

Anders,S. *et al.* (2015) HTSeq–a Python framework to work with high-throughput sequencing data. *Bioinformatics*, **31**, 166–169.

Bacher,R. and Kendziorski,C. (2016) Design and computational analysis of single-cell RNA-sequencing experiments. *Genome Biol.*, **17**, 63.

Bolger,A.M. *et al.* (2014) Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*, **30**, 2114–2120.

Borel,C. *et al.* (2015) Biased allelic expression in human primary fibroblast single cells. *Am. J. Hum. Genet.*, **96**, 70–80.

Buettner,F. *et al.* (2015) Computational analysis of cell-to-cell heterogeneity in single-cell RNA-sequencing data reveals hidden subpopulations of cells. *Nat. Biotechnol.*, **33**, 155–160.

Castel,S.E. *et al.* (2015) Tools and best practices for data processing in allelic expression analysis. *Genome Biol.*, **16**, 195.

Chen,G. *et al.* (2019) Single-cell RNA-seq technologies and related computational data analysis. *Front. Genet.*, **10**, 317.

Deng,Q. *et al*. (2014) Single-cell RNA-seq reveals dynamic, random monoallelic gene expression in mammalian cells. *Science*, **343**, 193–196.

Dobin,A. and Gingeras,T.R. (2015) Mapping RNA-seq reads with STAR. *Curr. Protoc. Bioinformatics*, **51**, 11.14.1–11.14.19.

Fan,H.C. *et al*. (2015) Expression profiling. Combinatorial labeling of single cells for gene expression cytometry. *Science*, **347**, 1258367.

Garieri,M. *et al*. (2018) Extensive cellular heterogeneity of X inactivation revealed by single-cell allele-specific expression in human fibroblasts. *Proc. Natl. Acad. Sci. USA*, **115**, 13015–13020.

Haque,A. *et al*. (2017) A practical guide to single-cell RNA-sequencing for biomedical research and clinical applications. *Genome Med*., **9**, 75.

Hashimshony,T. *et al*. (2016) CEL-Seq2: sensitive highly-multiplexed single--cell RNA-Seq. *Genome Biol*., **17**, 77.

Ilicic,T. *et al*. (2016) Classification of low quality cells from single-cell RNA-seq data. *Genome Biol*., **17**, 29.

Jiang,Y. *et al*. (2017) SCALE: modeling allele-specific gene expression by single-cell RNA sequencing. *Genome Biol*., **18**, 74.

Kang,H.M. *et al*. (2018) Multiplexed droplet single-cell RNA-sequencing using natural genetic variation. *Nat. Biotechnol*., **36**, 89–94.

Kim,J.K. *et al*. (2015) Characterizing noise structure in single-cell RNA-seq distinguishes genuine from technical stochastic allelic expression. *Nat. Commun*., **6**, 8687.

Klein,A.M. *et al*. (2015) Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. *Cell*, **161**, 1187–1201.

Kolodziejczyk,A.A. *et al*. (2015) The technology and biology of single-cell RNA sequencing. *Mol. Cell*, **58**, 610–620.

Lan,F. *et al*. (2017) Single-cell genome sequencing at ultra-high-throughput with microfluidic droplet barcoding. *Nat. Biotechnol*., **35**, 640–646.

Larsson,A.J.M. *et al*. (2019) Genomic encoding of transcriptional burst kinetics. *Nature*, **565**, 251–254.

Lun,A.T. *et al*. (2016a) Pooling across cells to normalize single-cell RNA sequencing data with many zero counts. *Genome Biol*., **17**, 75.

Lun,A.T. *et al*. (2016b) A step-by-step workflow for low-level analysis of single-cell RNA-seq data with Bioconductor. *F1000Res*., **5**, 2122.

Macaulay,I.C. *et al*. (2017) Single-cell multiomics: multiple measurements from single cells. *Trends Genet*., **33**, 155–168.

McCarthy,D.J. *et al*. (2017) Scater: pre-processing, quality control, normalization and visualization of single-cell RNA-seq data in R. *Bioinformatics*, **33**, 1179–1186.

McGinnis,C.S. *et al*. (2019a) DoubletFinder: doublet detection in single-cell RNA sequencing data using artificial nearest neighbors. *Cell Syst*., **8**, 329–337.

McGinnis,C.S. *et al*. (2019b) MULTI-seq: sample multiplexing for single-cell RNA sequencing using lipid-tagged indices. *Nat. Methods*, **16**, 619–626.

Pezzotti,N. *et al*. (2017) Approximated and user steerable tSNE for progressive visual analytics. *IEEE Trans. Vis. Comput. Graph*., **23**, 1739–1752.

Picelli,S. *et al*. (2014) Full-length RNA-seq from single cells using Smart-seq2. *Nat. Protoc*., **9**, 171–181.

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

Risso,D. *et al*. (2018) A general and flexible method for signal extraction from single-cell RNA-seq data. *Nat. Commun*., **9**, 284.

Sheng,K. *et al*. (2017) Effective detection of variation in single-cell transcriptomes using MATQ-seq. *Nat. Methods*, **14**, 267–270.

Sheng,K. and Zong,C. (2019) Single-cell RNA-seq by multiple annealing and tailing-based quantitative single-cell RNA-seq (MATQ-Seq). *Methods Mol. Biol*., **1979**, 57–71.

Stegle,O. *et al*. (2015) Computational and analytical challenges in single-cell transcriptomics. *Nat. Rev. Genet*., **16**, 133–145.

Stoeckius,M. *et al*. (2018) Cell hashing with barcoded antibodies enables multiplexing and doublet detection for single cell genomics. *Genome Biol*., **19**, 224.

Tang,F. *et al*. (2011) Deterministic and stochastic allele specific gene expression in single mouse blastomeres. *PLoS One*, **6**, e21208.

Tukiainen,T. *et al*.; GTEx Consortium. (2017) Landscape of X chromosome inactivation across human tissues. *Nature*, **550**, 244–248.

Usoskin,D. *et al*. (2015) Unbiased classification of sensory neuron types by large-scale single-cell RNA sequencing. *Nat. Neurosci*., **18**, 145–153.

Van der Auwera,G.A. *et al*. (2013) From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline. *Curr. Protoc. Bioinformatics*, **43**, 11.10.1–11.10.33.

Villani,A.C. *et al*. (2017) Single-cell RNA-seq reveals new types of human blood dendritic cells, monocytes, and progenitors. *Science*, **356**, eaah4573.

Wagner,J.M. *et al*. (2018) A comparative analysis of single cell and droplet-based FACS for improving production phenotypes: riboflavin overproduction in Yarrowia lipolytica. *Metab. Eng*., **47**, 346–356.

Wainer-Katsir,K. and Linial,M. (2019) Human genes escaping X-inactivation revealed by single cell expression data. *BMC Genomics*, **20**, 201.

Wolock,S.L. *et al*. (2019) Scrublet: computational identification of cell doublets in single-cell transcriptomic data. *Cell Syst*., **8**, 281–291.

Xin,Y. *et al*. (2016) Use of the Fluidigm C1 platform for RNA sequencing of single mouse pancreatic islet cells. *Proc. Natl. Acad. Sci. USA*, **113**, 3293–3298.

Zeisel,A. *et al*. (2015) Brain structure. Cell types in the mouse cortex and hippocampus revealed by single-cell RNA-seq. *Science*, **347**, 1138–1142.

Zhang,X. *et al*. (2019) Comparative analysis of droplet-based ultra-high--throughput single-cell RNA-seq systems. *Mol. Cell*., **73**, 130–142.

Zheng,G.X. *et al*. (2017) Massively parallel digital transcriptional profiling of single cells. *Nat. Commun*., **8**, 14049.

Zilionis,R. *et al*. (2017) Single-cell barcoding and sequencing using droplet microfluidics. *Nat. Protoc*., **12**, 44–73.