Genome analysis

Destin: toolkit for single-cell analysis of chromatin accessibility

Eugene Urrutia 1, Li Chen 2, Haibo Zhou 1 and Yuchao Jiang 1,3,4,*

1Department of Bioinformatics, Gillings School of Global Public Health, University of North Carolina, Chapel Hill, NC 27599, USA.
2Department of Health Outcomes Research and Policy, Harrison School of Pharmacy, Auburn University, Auburn, AL 36849, USA.
3Department of Genetics, School of Medicine, University of North Carolina, Chapel Hill, NC 27599, USA.
4Lineberger Comprehensive Cancer Center, University of North Carolina, Chapel Hill, NC 27599, USA.

To whom correspondence should be addressed.

Abstract

Summary: Single-cell assay of transposase-accessible chromatin followed by sequencing (scATAC-seq) is an emerging new technology for the study of gene regulation with single-cell resolution. The data from scATAC-seq are unique - sparse, binary, and highly variable even within the same cell type. As such, neither methods developed for bulk ATAC-seq nor single-cell RNA-seq data are appropriate. Here, we present Destin, a bioinformatic and statistical framework for comprehensive scATAC-seq data analysis. We evaluated performance of Destin using downsampled bulk ATAC-seq data of purified samples and scATAC-seq data from seven diverse experiments. Compared to existing methods, Destin was shown to outperform across all data sets and platforms. For demonstration, we further applied Destin to 2,088 adult mouse forebrain cells and identified cell type-specific association of previously reported schizophrenia GWAS loci.

Availability: Destin toolkit is freely available as an R package at https://github.com/urrutiag/destin.

1 Introduction

Single-cell assay of transposase-accessible chromatin followed by sequencing (scATAC-seq) is an emerging new technology for the study of gene regulation with single-cell resolution. Unlike conventional regulomics technologies, scATAC-seq measures chromatin accessibility within each individual cell, which circumvents the averaging artifacts associated with traditional bulk population data, yielding new insights into epigenetic regulation at the cellular level. Technologically, Buenrostro et al. (2015) adapted the bulk ATAC-seq technology to single cells, utilizing microfluidic device to physically isolate single cells. Cusanovich et al. (2015) adopted a two-step “split-and-pool” strategy, where cells undergo several rounds of barcoding procedures to be uniquely labeled. Recently, Preisol et al. (2018) developed single nucleus ATAC-seq (snATAC-seq), adapting the two-step combinatorial indexing strategy to frozen tissues. The Chromium Single Cell ATAC Solution by 10X Genomics can further profile chromatin accessibility across 500 to 10,000 nuclei in parallel. Refer to Supplementary Table 1 for a summary of existing platforms and technologies.

The data from scATAC-seq are unique - sparse, binary, noisy with biases and artifacts, and highly variable even within cell types. Supplementary Fig. 1 shows a snapshot of both single-cell and bulk-tissue chromatin accessibility within a 800kb region from chromosome 1, using human monocyte cells and purified bulk samples (Corces et al., 2016), respectively. ScATAC-seq data is sparse and noisy, and the signals show low similarity across cells. On the contrary, bulk signals are highly conserved across the five purified samples. Total depth of coverage in single cells is also greatly reduced, several orders of magnitude less than bulk. Additionally, since most of the genome has only two copies in a cell and the transposase can cleave and add adaptors only once per copy, for each open chromatin region, at most two scorable fragments can be generated and equivalently at most two reads per locus can be obtained after removing PCR duplicates. As such, scATAC-seq data is also highly binary, indicating an open/closed status. Because of the aforementioned uniqueness of scATAC-seq data, neither methods developed for bulk ATAC-seq nor single-cell RNA-seq (scRNA-seq) can be directly applied.

Several cell type clustering methods for scATAC-seq have been proposed. ScAsAT (Baker et al., 2017) first performs dimension reduction on cell-by-cell Jaccard distance matrix, followed by t-SNE (Maaten and Hinton, 2008) and k-methods for clustering. scABC (Zamanighomi et al., 2018) begins by clustering cell types using weighted k-methods, up-weighting cells with higher sequencing depth. To address the sparsity and noise observed in scATAC-seq and to reduce dimension, SCRAT (Ji et al., 2017) and chromVar (Schep et al., 2017) aggregate scATAC-seq read counts across biological features such as transcription factor binding motifs, DNase I hypersensitivity sites (DHSs), genes, or gene sets of interest. This is followed by dimension reduction and clustering.

Here, we propose Destin (Detection of cell-type specific difference in chromatin accessibility), a bioinformatic and statistical framework for comprehensive scATAC-seq data analysis. For cell type clustering, instead of aggregating peaks based on genomic annotations, Destin adopts weighted principal component analysis (PCA), with peak-specific weights calculated based on the distances to transcription start sites (TSSs) as well as the relative frequency of chromatin accessibility peaks based on a broad range of reference experiments. The weights, together with the number of clusters, and the number of principle components, are cast as tuning parameters and are determined based on the likelihood calculated from a post-classification multinomial model. Destin is evaluated on scATAC-seq data of 5,800 cells from seven experiments and is benchmarked against existing methods. We show that Destin outperforms the other methods across different data sets and platforms. As a proof of concept, we demonstrate Destin on a scATAC-seq data set of 2,088 adult mouse forebrain cells and identify cell type-specific association of previously reported GWAS loci for schizophrenia.

2 Materials and methods

Destin begins with a bioinformatic pipeline to preprocess raw sequencing files and follows with statistical analysis for cell type clustering and cell type enrichment of previously reported GWAS loci. Specifically, the bioinformatic pipeline includes...
demultiplexing (for platforms with cellular barcodes), trimming adapters, mapping reads, filtering blacklisted regions, calling peaks in pseudo-bulk samples as aggregates of single cells. After a further quality control procedure, this results in a peak-by-cell chromatin accessibility matrix. Refer to Supplementary Materials for details on bioinformatic analysis.

Destin utilizes biological annotation in order to prioritize chromatin accessible regions that are more informative to distinguish cell types. Specifically, Corces et al. (2016) showed that distal regulatory elements (e.g., enhancers) provide stronger ability for clustering than do proximal elements (e.g., promoters), and thus, for analysis, focused on chromatin accessible regions 1kb upstream from the TSS. Similarly, Preissl et al. (2018) focused on chromatin accessible regions outside a 2kb window from the TSS. However, Corces et al. (2016) also showed that there is a predictive value in the promoter region, while (Zamanighomi et al., 2018) further identified cluster-specific open promoters that distinguish expression. As such, Destin retains the "cell type specificity" for each gene, based on gene annotations of chromatin accessible regions and clustering results by Destin. The cell type specificity was then tested for association with GWAS p-values mapped to each gene from three psychiatric studies of schizophrenia (Ripke et al., 2014), major depressive disorder (Wray et al., 2018), and attention deficit hyperactivity disorder (Demontis et al., 2017). We discovered significant association between schizophrenia GWAS loci and inhibitory neuron 1 by two independent methods MAGMA (de Leeuw et al., 2015) and eHCWC (Skene and Grant, 2016).

Funding
This work has been supported by the National Institutes of Health (NIH) grant P01 CA142538, R15 GM118102, and a developmental award 2017TT109 from the UNC Lineberger Comprehensive Cancer Center (2017TT109) to YJ, and the NIH Ruth L. Kirschstein NRSA T32 ES007018 to EU. Conflict of Interest: none declared.

References
Baker, S. M., et al. (2017). Classifying cells with scasat-a tool to analyse single-cell atac-seq. bioRxiv, page 227397.
Buenrostro, J. D., et al. (2015). Single-cell chromatin accessibility reveals principles of regulatory variation. Nature, 522(7551), 421.
Corces, M. R., et al. (2016). Lineage-specific and single-cell chromatin accessibility charts human hematopoiesis and leukemia evolution. Nat. genet., 48(10), 1193.
Cusanovich, D. A., et al. (2015). Multiplex single-cell profiling of chromatin accessibility by combinatorial cellular indexing. Science, 348(6237), 918–914.
De Leuven, C. A., et al. (2015). Mamg generalized gene-set analysis of gwas data. PLoS comput. biol., 11(4), e1004219.
Demontis, D., et al. (2017). Discovery of the first genome-wide significant risk loci for adhd. BioRxiv, page 145861.
Ji, Z., et al. (2017). Single-cell gromlese data analysis by scat. Bioinformatic, 33(18), 2930–2932.
Maaten, L. v. d., et al. (2008). Visualizing data using t-stee. J. mach. learn. res., 9(Nov), 2579–2605.
Preissl, S., et al. (2014). Single-nucleus analysis of accessible chromatin in developing mouse forebrain reveals cell-type-specific transcriptional regulation. Technical report, Nature Publishing Group.
Ripke, S., et al. (2014). Biological insights from 108 schizophrenia-associated genetic loci. Nature, 511(7510), 421.
Scher, A. N., et al. (2017). Chromvar: inferring transcription-factor-associated accessibility from single-cell epigenomic data. Nat. methods, 14(6), 975.
Skene, N. G., et al. (2016). Identification of vulnerable cell types in major brain disorders using single cell transcriptionomes and expression weighted cell type enrichment. Front. neurosci., 10, 16.
Skene, N. G., et al. (2018). Genetic identification of brain cell types underlying schizophrenia. Nat. genet., page 1.
Wray, N. R., et al. (2018). Genome-wide association analyses identify 44 risk variants and refine the genetic architecture of major depression. Nat. genet., 50(5), 668.
Zamanighomi, M., et al. (2018). Unsupervised clustering and epigenetic classification of single cells. Nat. commun., 9(1), 2410.

Table 1. Cell type enrichment for previously reported GWAS loci.Bonferroni corrected p-values are calculated by MAGMA and eHCWC for association of mouse forebrain cell types with previously reported GWAS loci for schizophrenia (SCZ), major depressive disorder (MDD), and attention deficit hyperactivity disorder (ADHD).

| Cell Type | MAGMA p-value | eHCWC p-value |
|-----------|----------------|----------------|
| Inhibitory neuron 1 | 0.0087 | 0.016 |
| Inhibitory neuron 2 | 0.434 | 0.204 |
| Neuron 1 | 0.311 | 0.774 |
| Neuron 2 | 0.871 | 1 |
| Oligodendrocyte 1 | 0.03 | 0.03 |

3 Results
We benchmarked Destin against three existing scATAC-seq methods: scABC, ScAtAT and chromVar. The online GUI by SCRAT takes bam files as input and cannot handle datasets with large number of cells. We began our benchmarks by downsampling bulk ATAC-seq data of purified bulk samples from Corces et al. (2016). We varied the number of cell types (2, 4, 6, or 8) and downsampled 50 cells per cell type. Median read depth per cell was also varied, ranging from 3,000 (combinatorial barcode indexing) to 70,000 (Fluidigm C1), with mean cluster purity shown in Fig 1A. Our results show that performance for all methods increased with number of reads and that Destin outperformed all other methods in almost every scenario. Next, we benchmarked Destin against the other methods using the seven full single cell ATAC-seq data sets (Supplementary Table 2). Destin performed as well as or better than all other methods in terms of cluster purity across all data sets (Fig 1B), including cluster purity at nearly 100% in five out of seven data sets. We further applied Destin to scATAC-seq data of 2,088 cells from adult mouse forebrain Preissl et al. (2018). Destin’s cell type clustering results, by its default, differ from the results by the original publication – Destin clustered together both the three excitatory neuron subtypes and the related microglia and astrocytes, the latter of which we refer to as neuroglia (Supplementary Fig. 7A). Notably, the subclusters can be resolved by increasing the number of clusters (Supplementary Fig. 7B). A key application of single cellomics is to identify specific cell types that are associated with disease. Compared to scRNA-seq, the scope is broadened to outside the gene bodies to include, e.g., promoters and enhancers. In a similar fashion to Skene et al. (2018), we determined the "cell type specificity" for each gene, based on gene annotations of chromatin accessible regions and clustering results by Destin. The cell type specificity was then tested for association with GWAS p-values mapped to each gene from three psychiatric studies of schizophrenia (Ripke et al., 2014), major depressive disorder (Wray et al., 2018), and attention deficit hyperactivity disorder (Demontis et al., 2017). We discovered significant association between schizophrenia GWAS loci and inhibitory neuron 1 by two independent methods MAGMA (de Leeuw et al., 2015) and eHCWC (Skene and Grant, 2016).