Published in final edited form as:
Nat Methods. 2017 October ; 14(10): 975–978. doi:10.1038/nmeth.4401.

chromVAR: Inferring transcription factor-associated accessibility from single-cell epigenomic data

Alicia N. Schep[1,2], **Beijing Wu**[1,2], **Jason D. Buenrostro**[3,4,*], and **William J. Greenleaf**[1,2,5,*]

[1]Department of Genetics, Stanford University School of Medicine, Stanford, CA 94305, USA

[2]Center for Personal Dynamic Regulomes, Stanford University, Stanford, CA 94305

[3]Broad Institute of MIT and Harvard, Cambridge, MA 02142, USA

[4]Harvard Society of Fellows, Harvard University, Cambridge, MA 02138, USA

[5]Department of Applied Physics, Stanford University, Stanford, CA 94025, USA

Abstract

Single cell ATAC-seq (scATAC) yields sparse data that makes application of conventional analysis approaches challenging. We developed chromVAR, an R package for analyzing sparse chromatin accessibility data by estimating gain or loss of accessibility within peaks sharing the same motif or annotation while controlling for technical biases. chromVAR enables accurate clustering of scATAC-seq profiles and enables characterization of known and *de novo* sequence motifs associated with variation in chromatin accessibility.

Main Text

Transcription factor binding to regulatory DNA sequences controls the activity of *cis*-regulatory elements, which modulate gene expression programs that define cell phenotype. Assays for probing chromatin accessibility have enabled the discovery of *cis*-regulatory elements and *trans*-acting factors across different cell states and types that lead to functional changes in gene expression[1]. Concurrently, single cell genomic and transcriptomic methods have enabled unbiased *de novo* deconvolution of dynamic or diverse cellular populations[2,3]. Recently-developed assays for measuring chromatin accessibility within single cells[4–6] promise to enable the identification of causative *cis*- and *trans*- regulators that bring about these diverse cellular phenotypes.

However, the exceedingly sparse nature of single cell epigenomic data sets present unique and significant computational challenges. All single-cell epigenomic methods are

[*]Authors for correspondence: jbuen@broadinstitute.org or wjg@stanford.edu.

Competing Interests
J.D.B. and W.J.G. are listed as inventors on a patent for the ATAC- seq method. W.J.G. is a scientific cofounder of Epinomics.

Author Contributions
A.N.S, J.D.B, and W.J.G conceived the project and wrote the manuscript. A.N.S. wrote the chromVAR R package and performed the analyses with input from J.D.B and W.J.G. B.W. generated the scATAC-seq data.

intrinsically sparse, as the total potential signal at a genomic locus is fundamentally limited by the copy number of DNA, thus generating 0, 1 or 2 reads from regulatory elements within a diploid genome. Methods developed for single cell RNA-seq have shown that measuring the dispersion of gene sets, such as Gene Ontology or co-expression modules, rather than individual genes can be a powerful approach for analyzing sparse data[7]. In this vein, and building on previous work from our group and others[4,8,9], we have developed *chromVAR*, a versatile R package for analyzing sparse chromatin accessibility data by measuring the gain or loss of chromatin accessibility within sets of genomic features while controlling for known technical biases in epigenomic data (Supplementary Software). We show that *chromVAR* can be used to identify transcription factor (TF) motifs that define different cell types and vary within populations, providing a unique analytical toolkit for analysis of sparse epigenomic data.

The chromVAR package takes as inputs 1) aligned sequencing reads, 2) chromatin accessibility peaks (derived from either data aggregated across cells or external resources), and 3) a set of chromatin features representing either motif position weight matrices (PWMs) or genomic annotations (Fig. 1a, Supplementary Fig. 1). For use as input (3) into chromVAR, we have curated a set of human and mouse PWMs from the cisBP database[10] that represent a diverse and comprehensive collection of known TF motifs. Alternately, user provided TF motifs or other types of genomic annotations, such as enhancer modules, ChIP-seq peaks, or GWAS disease annotation may be used. chromVAR may also be applied to a collection of kmers—DNA sequences of a specific length k—in order to perform an unbiased analysis of DNA sequence features that correlate with chromatin accessibility variation across the cells or samples.

chromVAR first computes a “raw accessibility deviation” for each motif and cell, representing the difference between the total number of fragments mapping to peaks containing the given motif and the total expected number of fragments based on the average of all input cells. This aggregation across peaks sharing a common motif yields a signal that is considerably less sparse than the signal within individual peaks, however this aggregation can also amplify technical biases between cells due to PCR amplification or variable Tn5 tagmentation conditions (Supplementary Note 1). These technical biases can lead to differences in the number of observed fragment counts between cells for a given peak set with distinct GC content or mean accessibility (Supplementary Fig. 2). To account for these technical confounders, “background” peak sets are created for each annotation, which comprise of an equal number of peaks matched for GC content and average accessibility (Supplementary Figs. 2–5; Supplementary Note 2). The raw accessibility deviations for these background peak sets are used to compute a bias corrected deviation and z-score for each annotation and cell, providing a bias-corrected differential measure describing the gain or loss of accessibility of a given genomic annotation relative to the average cell profile (**see methods**). These bias corrected deviations and z-scores can be used for a number of downstream analyses, including *de novo* clustering of cells and identification of key regulators that vary within and between different cell types. The chromVAR package includes a collection of tools for such downstream analysis, including an interactive web application for exploring the relationship between key TF motifs and clustering of cells

(Supplementary Figure 6). We have also incorporated tools for generating previously-described analyses characterizing the correlation and potential cooperativity between two TF binding sites within the same regulatory element, and computing chromatin variability across regions in *cis*[4].

To test the applicability of this computational workflow for single-cell analysis, we set out to measure the robustness of *chromVAR* outputs to data downsampling. To do this, we applied chromVAR to bulk ATAC-seq data from a deeply-sequenced set of hematopoietic cell types[8], and compared the results of the analysis for the data across various degrees of downsampling. We found that the TF motif deviations using 10^6 to 5×10^3 fragments per sample are highly correlated to those determined using the full data set (Fig. 1b, Supplementary Fig. 7). The clustering accuracy using the bias corrected deviations is also largely preserved after downsampling, and compares favorably to clustering using PCA or other peak-based approaches (Supplementary Fig. 7; **see methods**).

Importantly, chromVAR provides robust results for 10,000 fragments per cell, a typical number of fragments generated from a single-cell using scATAC-seq[4] (Supplementary Figure 7). By projecting the vector of bias corrected deviations from individual cells into two dimensions using tSNE[11], *chromVAR* enables the reconstruction of the major hematopoietic lineages using 10,000 fragments per sample. With this analytical framework, we can also visualize the TFs associated with significant chromatin accessibility within each simulated single cell epigenome, thereby correctly identifying known master regulators of hematopoiesis, including HOXA9, SPI1, TBX21, and GATA1[12–15] (Fig. 1c).

We next characterized chromVAR's ability to capture biologically relevant chromatin variability from single cell ATAC-seq data drawn from multiple distinct cell lines and human samples (Supplementary Fig. 8). Using tSNE with bias corrected deviations for motifs and 7mers, we clustered individual cells into distinct cell types and observe individual motifs that best distinguish each cell type (Fig. 2a). Notably, well defined, distinct clusters are formed in this tSNE projection when using the bias corrected deviations, but the clustering is confounded by technical biases when using raw deviations without the bias correction infrastructure. Importantly this approach for classifying cell types also compares favorably performing tSNE on the counts within peaks using a variety of approaches (Supplementary Fig. 9). Interestingly, we also observe that cells from acute myeloid leukemia (AML) patients cluster between lymphoid-primed multipotent progenitors (LMPPs), monocytes, and HL60 (an AML derived cancer cell line) cells. In this unsupervised analysis, we find that the AML leukemic stem cells are more similar to LMPPs, while the AML blasts are more similar to the monocytes. In addition, we also observe that patient 1 (AML blast 1) maintains a more stem-like state when compared to patient 2 (AML blast 2) as anticipated from alternate analyses of these cells[16]. By visualizing the cell-specific Z-scores layered on this projection, we identify putative TFs that may promote the stem-like versus differentiated leukemia phenotype; for example, the master-regulators of myeloid cell development SPI1 (PU.1) and CEBPA[17] appear as the most differential motifs between AML leukemic stem cells (LSCs) and blasts (Fig. 2b–c).

In addition to visualizing the similarity of cells, we inverted our tSNE analysis to visualize the similarity of motifs and kmers in their activity patterns across cells (Fig. 3a). In this visualization, motifs and kmers that have similar activity profiles across cells cluster together in the tSNE subspace, allowing the identification of major clusters representing several different TF families. Notably, different TFs within the same family (e.g. GATA1 and GATA2) often bind highly similar motifs, and therefore chromVAR alone cannot distinguish the causative regulator binding a particular TF motif. In the inverted tSNE visualization for motif and kmer similarity, most, but not all kmers cluster with a known motifs, suggesting k-mer analysis may enable *de novo* discovery of previously unannotated motifs.

By comparing the variation in chromatin accessibility across cells between highly similar kmers, we can identify critical bases associated with chromatin accessibility variation. For example, the “AGATAAG” kmer, which closely matches the GATA1 motif, is highly variable across single cells, but most kmers differing by one nucleotide share little or none of that variability (Fig. 3b, Supplementary Fig. 10). The mismatched kmer with the greatest correlated variability is “TGATAAG”, which is consistent with the weights of each nucleotide in the GATA1 motif. Similarly strong sequence specificity is seen across other variable motifs (Supplementary Fig. 10).

We can use these comparisons of variation between highly similar kmers to construct *de novo* motifs representing sequences associated with variation in chromatin accessibility. In brief, we start with highly variable “seed” kmers, and use the covariance between the seed kmer and kmers either differing by one mismatch or partially overlapping the seed kmer to assign weights to different nucleotide bases at each position of the motif model (Supplementary Fig. 11; **see methods**). Importantly, many *de novo* motifs assembled using this approach closely match known motifs (Figs. 3c–f, Supplementary Fig. 11). For motifs that do not closely match to a known TF, we confirmed that the constructed motifs were also associated with variation in DNase hypersensitivity between different samples represented in the Roadmap Epigenomics Project[18] (Supplementary Fig. 12), demonstrating that these *de novo* motifs are associated with chromatin accessibility variation in two distinct accessibility assays. To further validate the discovery of these putative *trans*-regulators we calculated aggregate TF “footprints”, a measure of the DNase or Tn5 cut density around the given motif, and found a diverse set of accessibility profiles (Supplementary Fig. 12). Interestingly, several of these motifs did not match canonical narrow (~20 bp) transcription factor footprints, but rather are associated with a large footprint (>20 bp) potentially indicative of larger regulatory complexes.

In summary, we envision that *chromVAR* will be broadly applicable to single-cell and bulk epigenomics data to provide an unbiased characterization of cell types and the *trans* regulators that define them. As such, we applied chromVAR to two bulk chromatin accessibility data sets[18,19] down-sampled to 10,000 fragments per sample and data from an alternate scATAC-seq approach and find chromVAR to generalize to these additional data (Supplementary Figs 13–15; Supplementary Note 3). As methods for measuring the epigenome in single-cells and bulk populations continue to improve in throughput and in quality, scalable analytical infrastructure is needed. Analysis workflows for ATAC or DNase-seq data often include the identification of motifs enriched in differentially accessible peaks, but such approaches scale poorly to comparisons across many sample types and require sufficient read depth per-locus to determine differential peak accessibility (Supplementary Note 4). In contrast, *chromVAR* analysis is highly robust to low sequencing depth and readily scales to hundreds or thousands of cells or samples. Budget-constrained researchers often face a trade-off between the number of samples to sequence and the sequencing depth for each sample; sparse sequencing analysis coupled with chromVAR analysis may enable new applications of “bulk” ATAC, DNase-seq or other epigenomic methods as large-scale screening tools. We also anticipate that *chromVAR* will enable additional downstream analyses of single cell chromatin accessibility data, as the reduction of dimensionality associated with vectors of bias corrected deviations provide a powerful input to existing algorithms for inferring inferring spatial and temporal relationships between cells.

Online Methods

chromVAR algorithm

Bias corrected deviations and z-scores—For each motif (or kmer or genomic annotation), a “raw accessibility deviation” for each cell or sample is computed that represents the difference in the total accessibility of peaks with that motif minus the expected count based on the accessibility profile across all cells, divided by that expected count (Figure S1). Using the matrix of fragment counts in peaks **X**, where $x_{i,j}$ represents the number of fragments from cell i in peak j, and the matrix of motif matches **M**, where $m_{k,j}$ is 1 if motif k is present in peak j. The total number of reads mapping to every peak containing motif k in cell i is given by $M * X^T$. For each peak, the expected number of fragments per cell E is computed as the fraction of all fragments across all cells mapping to that peak multiplied by the total number of fragments in peaks for that cell:

$$E = \frac{\sum_{i=1} x_{i,j}}{\sum_{j=1} \sum_{i=1} x_{i,j}} * \sum_{j=1} x_{i,j}$$

The expected number of fragments mapping to every peak containing motif k in cell i is then given by $M * E^T$, and the raw accessibility deviation Y by:

$$Y = \frac{M * X^T - M * E^T}{M * E^T}$$

For each motif or genomic annotation, background peak sets are sampled that match the set of peaks with the motif or genomic annotation in terms of the distribution of GC content and average accessibility. These background peak sets are determined by finding possible background peaks for each peak, as described in the next section. For each background iteration, we can represent the background peak assignments as a matrix $\boldsymbol{B}$ where $b_{j,j}$, is 1 if peak j has peak j' as it’s background peak and 0 otherwise. A background motif match matrix M′ is thus computed as $M' = M * B$, and a background raw deviation as:

$$Y' = \frac{(M*B)*X^T - (M*B)*E^T}{M*E^T}.$$

Y' is calculated for each background iteration, and these background deviations are used to compute a bias-corrected deviation as $Y - mean(Y')$. A deviation Z-score is computed by dividing the bias-corrected deviation by the standard deviation of the background raw deviations for each iteration:

$$\frac{Y - mean(Y')}{sd(Y')}.$$

Background peak selection—The state space of GC content and the log of the average accessibility of peaks is transformed by the Mahalanobis transformation in order to remove the correlation between the two variables. This transformed space is split into an even grid of bins with a specified number of divisions (50) along each axis evenly spaced between the minimum and maximum values. For a peak in a given bin j, the probability of selecting another peak x in bin i is given by:

$$P(x|x \in b_i) = \frac{f(d(i-j)|0, w)}{\rho_i}$$

Where f is the probability distribution function of the normal distribution with mean zero and standard deviation w (set to 0.01), and ρ is the number of peaks in the bin j.

Variability—The variability of a TF motif across samples or cells was determined by computing the standard deviation of the z-scores across the cells or samples. The expected value of this metric is one if the motif peak sets are no more variable than the background peak sets for that motif.

De novo motif assembly—As a measure of the shared variability in chromatin accessibility between a reference kmer (or motif) and other kmers (or motifs), we compute a normalized co-variance based on deviation z-scores. This normalized covariance is simply the covariance of the z-scores across each cell divided by the variance of the z-scores for the reference kmer (or motif).

For assembling *de novo* motifs, we start with the kmer associated with the greatest variability in chromatin accessibility across the cells as a "seed" kmer. We first find the distribution of the normalized covariances between that seed kmer and all other kmers with an edit distance from that seed kmer of at least 3; these values are used as a null distribution for testing the significance of the observed covariances for kmers with a single nucleotide mismatch using a Z-test. For each position along the kmer, the nucleotide of the seed kmer is given a weight of 1. Each alternate nucleotides is given a weight of zero if the p-value for the normalized covariance of the kmer with that mismatch is greater than 0.05; if the p-value is less than 0.05 the nucleotide is given a weight equal to the square of the normalized covariance. The weights for each base pair are then normalized to sum to 1. To further extend the *de novo* motif, we used kmers overlapping the seed kmer with an offset of 1 or 2 bases. For the two bases immediately outside the seed kmer, the weighting of each nucleotide is given by $x * y^2 + (1 - x)) * 0.25$, where y^2 is the square of the normalized covariance for the kmer with the given nucleotide offset (if significant at 0.05 and otherwise 0) and x is the maximum value of the normalized covariances for the four kmers (bounded by 0 and 1). For the bases offset by two from the seed kmer, the weighting is computed in the same way except that there are four possible kmers with a given nucleotide at that position that overlap the seed kmer; only the kmer with the maximum normalized covariance with the seed kmer is used (Figure S11).

Input data and pre-processing

ATAC-seq, scATAC, and DNase Data—In addition to the previously published data, we generated three new replicates of single-cell K562s (ATCC; validated using STR genotyping (Genetica DNA laboratories)) using the previously published protocol[4,8]. Bulk ATAC-seq and scATAC-seq data was aligned and filtered as described previously[4,8]. Uniformly processed DNase data was downloaded from the Roadmap Epigenomics Project Portal[18]. ATAC-seq data from Lavin et al. (2014) was obtained from GSE63341 and processed as follows: adapters were trimmed using Cutadapt[20], reads were aligned using Bowtie2[21], and filtered for mapping quality (mapq > 30). For the scATAC-seq data from the GM12878 and HEK293T mixture from the combinatorial indexing approach, a count matrix was obtained from GSM1647122.

Peaks—For the bulk data analysis, we obtained DNase hypersensitivity peaks from the Roadmap Epigenomics Project. MACS2[22] peaks for blood cells (Primary monocytes from peripheral blood, Primary B cells from peripheral blood, Primary T cells from peripheral blood, Primary Natural Killer cells from peripheral blood, Primary hematopoietic stem cells G-CSF-mobilized Female, Primary hematopoietic stem cells G-CSF-mobilized Male, and Monocytes-CD14+ RO01746 Cell line) were downloaded from the Epigenomics Roadmap Portal[18]. For the single cell ATAC-seq data, peaks were called for each cell line or type using MACS2 applied to the merged single cell ATAC-seq data. All peaks were re-sized to a uniform width of 500 bp, centered at the summit. For both the set of peak calls from the blood cells in Roadmap and the set of peak calls from the scATAC-seq data, peaks were combined by removing any peaks overlapping with a peak with greater signal. Peak width was chosen based on typical sizes of ATAC-seq peaks across a wide collection of experiments, although chromVAR is fairly robust to the exact size of the peaks used (Figure S5, Supplementary Note 2).

Motif collection—We curated Position Frequency Matrices from cisBP representing a total of 15,389 human motifs and 14,367 mouse motifs. To filter motifs to a representative subset, we first categorized motifs as high, medium or low quality, as is provided in the cisBP database. We then grouped all 870 unique human or 850 unique mouse TF regulators represented in the database and assigned these regulators to their most representative TF motif(s). To do this, we first iterated through each TF regulator to find all motifs associated with that regulator from the high-quality motif list. For these associated high quality motifs, we first computed a similarity matrix using the Pearson correlation of the motifs. To calculate the Pearson correlation between pair-wise motifs, the shorter motif was padded with an equal distribution of A,C,G,T. Then the Pearson correlation was calculated at every possible offset, and the maximum correlation of all offset comparisons was recorded. To select a representative subset of motifs for each TF regulator, we first found the motif correlated with the most other motifs at R>0.9. Treating that motif and all of the correlated motifs (R>0.9) as a group, we next found the motif with the greatest mean correlation to the other members of the group, and kept that motif as a representative motif for the TF. Motifs highly correlated with that chosen motif (R>0.9) were then discarded from further analysis, and the process was iterated until no motifs remained. We repeated the process using the medium and low-quality databases for TF regulators with no associated motifs in the high-quality database. The final curated motif database contains 1,764 human motifs and 1,346 mouse motifs representing 870 human and 850 mouse regulators. The resulting names are formatted as follows:: “ensemble ID”_”unique line number”_”common TF name”_”direct (D) or inferred (I)”_”number of similar motifs grouped”. These position frequency matrices were then converted into Position Weight Matrices (PWMS) by taking the log of the frequency after adding a 0.008 pseudocount and dividing by 0.25.

These PWMs were used for all analyses in main text figures. For Figures S2–5 and S13 a smaller set of motifs from the JASPAR CORE database 2016 were used[23]. For Figure S14, motifs downloaded from http://homer.ucsd.edu/homer/custom.motifs were used[24], and for Figure S15 motifs downloaded from http://compbio.mit.edu/encode-motifs/ were used[25] in order to use the same motifs as the original publication for those data sets.

Motif matching—The MOODS[26] C++ library (Version 1.9.3) was used for identifying peaks containing a motif match, using a p-value cutoff of 5×10^{-5}. As background frequencies we used the nucleotide frequencies across all peaks. We wrapped the MOODS library into an R package, *motifmatchr*, which enables fast determination of motif presence or positions within genomic regions. The package is available at www.github.com/GreenleafLab/motifmatchr and https://bioconductor.org/packages/devel/bioc/html/motifmatchr.html.

Analysis

Downsampling Analysis—To downsample a sample with X total fragments to a depth of Y total fragments, we use the fragment count matrix and for each fragment within a peak retained each fragment with probability Y/X. Thus the downsampled samples are equivalent to having approximately Y total fragments, but not precisely.

The set of peaks used for the analysis remained the same for each down-sampled data set, as the peaks used were from an external data source (Roadmap Epigenomics Project).

For clustering samples using chromVAR results, highly correlated motifs were first removed and then one minus the pearson correlation of the bias corrected deviations was used as the distance matrix for input into hierarchical clustering. For clustering samples using PCA, PCA was performed on the log of the fragment counts for all peaks after normalization for the total number of reads in peaks, and clustering was performed on the euclidean distance between the first five principal components. Hierarchical clustering was performed with complete linkage, and the resulting dendrogram was cut into 13 groups (the number of cell types). Clustering accuracy was measured using normalized mutual information [27].

Differential Accessibility and Variability—For determining differentially accessible motifs between AML LSC and blast cells, an unequal variances t-test (two-sided) was used on the bias corrected deviations. For determining differential variability, a Brown–Forsythe test was used on the deviation z-scores.

Sample similarity tSNE—For performing sample similarity tSNE, highly correlated motifs or kmers as well as motifs or kmers with variability below a certain threshold (1.5) were first removed from the bias corrected deviations matrix. The transpose of that matrix was then used as input to the Rtsne package[28], with a perplexity parameter of 8 used for the down-sampled bulk hematopoiesis data and 25 for the single cell ATAC-seq data.

Motif and kmer similarity tSNE—For performing motif similarity tSNE, motifs or kmers with variability below a certain threshold (1.5) were first removed from the bias corrected deviations matrix, which was then used as input to the Rtsne package [28] with perplexity parameter set to 15.

Motif Similarity Scores—To score the similarity between a de novo motif and the most similar known motif, we first computed the normalized Euclidean distance between the de novo motif and all the known motifs in our collection using the optimal local alignment with at least five overlapping bases. We then selected the known motif with the lowest distance as the closest match. The similarity score was computed as the negative of the Z-score for this distance using the distribution of distances for all the motifs in the collection.

Software Availability

The chromVAR R package is freely available under the MIT license at www.github.com/GreenleafLab/chromVAR and as Supplementary Software. The motifmatchr R package is freely available under a GPL-3 license is available at www.github.com/GreenleafLab/motifmatchr and as supplementary software.

Data Availability

The additional K562 scATAC-seq data have been deposited at GEO with accession number GSE99172. Previously published single cell ATAC-seq data are available from GSE74310 and GSE65360. Bulk hematopeisis ATAC-seq data are available at GSE74912. Macrophage bulk ATAC-seq data was obtained from GSE63341, combinatorial scATAC-seq from GSM1647122, and Roadmap Epigenomics data from the Roadmap Epigenomics Portal (http://egg2.wustl.edu/roadmap/web_portal/).

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This work was supported by National Institutes of Health (NIH) P50HG007735 (to W.J.G.), U19AI057266 (to W.J.G.), the Rita Allen Foundation (to W.J.G.) and the Baxter Foundation Faculty Scholar Grant and the Human Frontiers Science Program (to W.J.G). J.D.B. acknowledges support from the Harvard Society of Fellows and Broad Institute Fellowship. ANS acknowledges support from the National Science Foundation (NSF) GRFP (DGE-114747). We thank Caleb Lareau for valuable suggestions for improvements on the package as well as members of Greenleaf and Buenrostro labs for useful discussions.

References

1. Thurman RE, et al. The accessible chromatin landscape of the human genome. Nature. 2012; 489:75–82. [PubMed: 22955617]
2. Tang F, et al. Tracing the derivation of embryonic stem cells from the inner cell mass by single-cell RNA-Seq analysis. Cell Stem Cell. 2010; 6:468–478. [PubMed: 20452321]
3. Jaitin DA, et al. Massively parallel single-cell RNA-seq for marker-free decomposition of tissues into cell types. Science. 2014; 343:776–779. [PubMed: 24531970]
4. Buenrostro JD, et al. Single-cell chromatin accessibility reveals principles of regulatory variation. Nature. 2015; 523:486–490. [PubMed: 26083756]
5. Jin W, et al. Genome-wide detection of DNase I hypersensitive sites in single cells and FFPE tissue samples. Nature. 2015; doi: 10.1038/nature15740
6. Cusanovich DA, et al. Multiplex single cell profiling of chromatin accessibility by combinatorial cellular indexing. Science. 2015; 348:910–914. [PubMed: 25953818]
7. Fan J, et al. Characterizing transcriptional heterogeneity through pathway and gene set overdispersion analysis. Nat Methods. 2016; 13:241–244. [PubMed: 26780092]
8. Ryan Corces M, et al. Lineage-specific and single-cell chromatin accessibility charts human hematopoiesis and leukemia evolution. Nat Genet. 2016; 48:1193–1203. [PubMed: 27526324]
9. Farlik M, et al. Single-cell DNA methylome sequencing and bioinformatic inference of epigenomic cell-state dynamics. Cell Rep. 2015; 10:1386–1397. [PubMed: 25732828]
10. Weirauch MT, et al. Determination and inference of eukaryotic transcription factor sequence specificity. Cell. 2014; 158:1431–1443. [PubMed: 25215497]
11. van der, Maaten L., Hinton, G. Visualizing Data using t-SNE. J Mach Learn Res. 2008; 9:2579–2605.
12. Fujiwara Y, Browne CP, Cunniff K, Goff SC, Orkin SH. Arrested development of embryonic red cell precursors in mouse embryos lacking transcription factor GATA-1. Proc Natl Acad Sci U S A. 1996; 93:12355–12358. [PubMed: 8901585]
13. Ramos-Mejía V, et al. HOXA9 promotes hematopoietic commitment of human embryonic stem cells. Blood. 2014; 124:3065–3075. [PubMed: 25185710]
14. Nerlov C, Graf T. PU. 1 induces myeloid lineage commitment in multipotent hematopoietic progenitors. Genes Dev. 1998; 12:2403–2412. [PubMed: 9694804]
15. Gordon SM, et al. The transcription factors T-bet and Eomes control key checkpoints of natural killer cell maturation. Immunity. 2012; 36:55–67. [PubMed: 22261438]
16. Goardon N, et al. Coexistence of LMPP-like and GMP-like leukemia stem cells in acute myeloid leukemia. Cancer Cell. 2011; 19:138–152. [PubMed: 21251617]
17. Zhang P, et al. Enhancement of hematopoietic stem cell repopulating capacity and self-renewal in the absence of the transcription factor C/EBP alpha. Immunity. 2004; 21:853–863. [PubMed: 15589173]
18. Roadmap Epigenomics Consortium et al. Integrative analysis of 111 reference human epigenomes. Nature. 2015; 518:317–330. [PubMed: 25693563]
19. Lavin Y, et al. Tissue-resident macrophage enhancer landscapes are shaped by the local microenvironment. Cell. 2014; 159:1312–1326. [PubMed: 25480296]
20. Martin M. Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet.journal. 2011; 17:10–12.

21. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nat Methods. 2012; 9:357–359. [PubMed: 22388286]

22. Zhang Y, et al. Model-based analysis of ChIP-Seq (MACS). Genome Biol. 2008; 9:R137. [PubMed: 18798982]

23. Mathelier A, et al. JASPAR 2016: a major expansion and update of the open-access database of transcription factor binding profiles. Nucleic Acids Res. 2015; doi: 10.1093/nar/gkv1176

24. Heinz S, et al. Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. Mol Cell. 2010; 38:576–589. [PubMed: 20513432]

25. Kheradpour P, Kellis M. Systematic discovery and characterization of regulatory motifs in ENCODE TF binding experiments. Nucleic Acids Res. 2014; 42:2976–2987. [PubMed: 24335146]

26. Korhonen J, Martinmäki P, Pizzi C, Rastas P, Ukkonen E. MOODS: fast search for position weight matrix matches in DNA sequences. Bioinformatics. 2009; 25:3181–3182. [PubMed: 19773334]

27. Danon L, Díaz-Guilera A, Duch J, Arenas A. Comparing community structure identification. J Stat Mech. 2005; 2005:P09008–P09008.

28. Krijthe, J. Rtsne: T-Distributed Stochastic Neighbor Embedding using Barnes-Hut Implementation. R package version 0. 10. 2015. URLhttp://CRAN.R-project.org/package=Rtsne

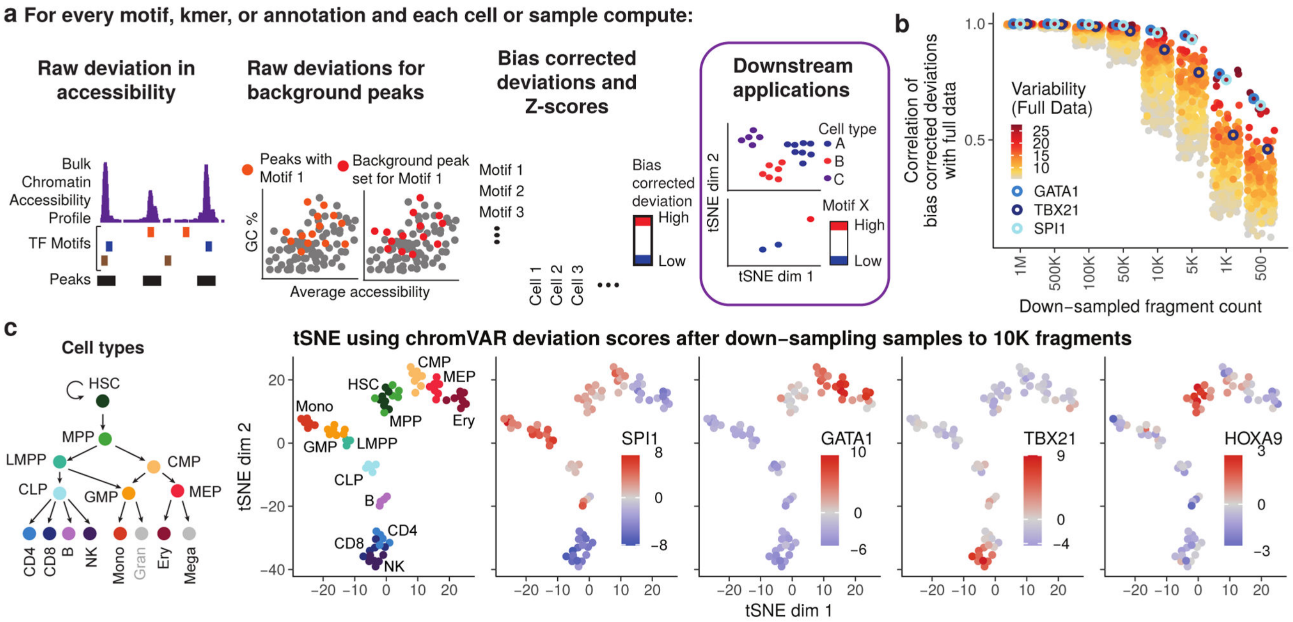

Figure 1. *chromVAR* enables interpretable analysis of sparse chromatin accessibility data

(a) Schematic illustrating how chromVAR uses aggregation of accessibility across peaks sharing a common feature (e.g. a motif) with bias correction to generate scores for each cell or sample that can be used for downstream analysis (b) Pearson correlation of bias corrected deviations for 77 samples from different hematopoietic populations before and after downs down-sampling total sequencing reads from full data. Each point shows the correlation for a different motif. The top 20% most variable motifs are shown. Three of the most variable motifs are highlighted. (c) tSNE visualization of different samples using normalized deviations calculated from data down-sampled to 10,000 fragments per sample. In the first panel, cells are colored by cell type, and in other panels cells are colored by the deviations score for different motifs.

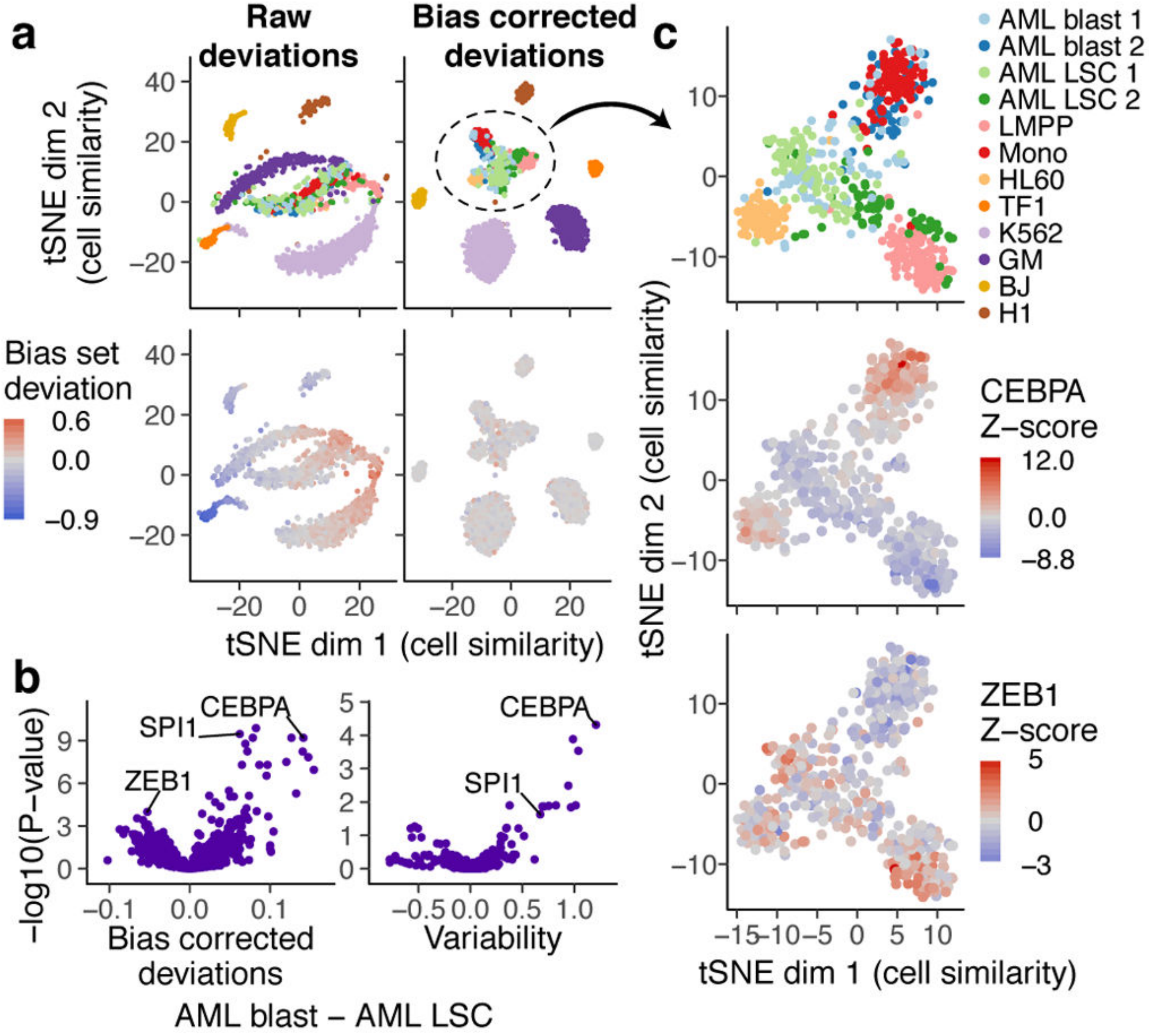

Figure 2. *chromVAR* enables clustering of single cell populations and interpretation of motifs underlying chromatin accessibility variation in single cells

(a) tSNE visualization of similarity of 1561 single cells based on chromVAR raw (left) or bias corrected deviations (right) for motifs and 7mers (see methods). In top panels, points are colored by cell type and in bottom panels points are colored by raw (left) or bias corrected (right) calculated deviations for a set of random peaks with high GC content and high average accessibility (the bias set). (b) Volcano plot showing the mean difference in bias corrected accessibility deviations (left) and variability (right) for each motif between the AML blast (n = 122) and LSC cells (n = 144) versus the $-\log_{10}$(P-value) for that difference. (c) tSNE with bias corrected deviations for AML blast and LSC, monocyte, LMPP, and HL60 cells (n = 509). In top panel, points are colored by cell type, and in other panels points are colored by deviation Z-scores for CEBPA and ZEB1 respectively.

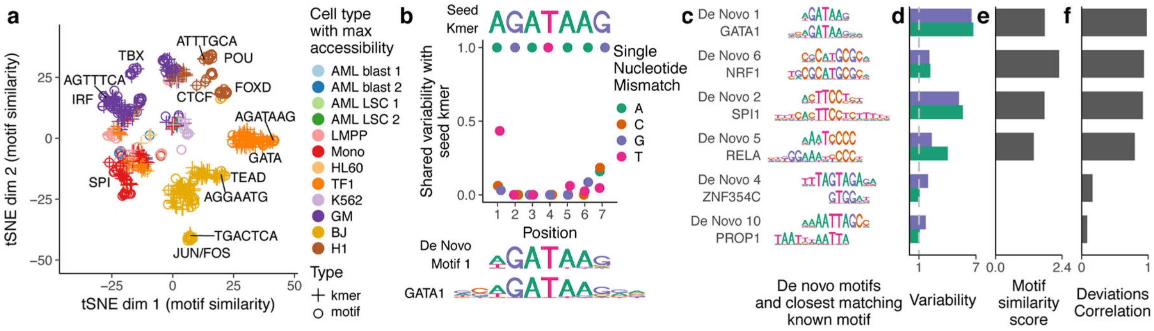

Figure 3. *chromVAR* identifies *de novo* motifs associated with chromatin accessibility variation in single cells

(a) tSNE visualization of similarity between motifs and kmers based on the vector of normalized deviations across different cells. Labels highlight predominant families of motifs within a cluster and example kmers (b) For the seed kmer "AGATAAG", the shared variability of k-mers with one mismatch from the seed kmer. The shared variability is defined as the square of the covariance of the deviation z-scores for the two kmers divided by the variance of the seed kmer for covariances greater than zero, and zero otherwise. These shared variabilities were used to assemble a *de novo* motif, shown under the plot along with the GATA1 motif. (c) Example *de novo* motifs assembled by chromVAR using deviations scores for 7-mers, along with the closest matching known motif below it. (d) Variability for both the *de novo* motif and the known motif for each pair in panel (c). (e) Motif similarity score (see methods) between the *de novo* motif and the known motifs in (c) (f) The Pearson correlation between the normalized deviations of the *de novo* motif and the known motif for each pair in (c).