doi:10.1038/nature14590

Supplementary Materials & Methods

1 - Equation glossary

2 - Experimental Methods

2.1 - Cell culture

K562 (ATCC) chronic myeloid leukemia cells were maintained in Iscove's modified Dulbecco's medium (IMDM) containing 10% FBS (HyClone, Thermo Scientific) and 1% Penicillin Streptomycin (Pen/Strep). K562s were validated using STR genotyping (Genetica DNA laboratories). GM12878 (ATCC) lymphoblastoid cells were grown in RPMI 1640 with 2 mM L-glutamine, 15% FBS and 1% Pen/Strep. The mouse EML hematopoietic cells (ATCC) were grown in IMDM containing 20% horse serum, 2 mM L-Glutamine, 1% Pen/Strep and 100 ng/mL recombinant SCF (Peprotech). The erythroleukemia cell line TF-1 (kind gift from the Majeti lab, Stanford) was maintained in RPMI 1640 with 10% FBS, 1% Pen/Strep and 2 ng/mL rhGM-CSF (Peprotech). The promyelocytic leukemia cells HL-60 (ATCC) were grown in IMDM containing 20% FBS and 1% Pen/Strep. The H1 human embryonic stem cells were obtained from WiCell and grown in chemically defined mTeSR1 medium (STEMCELL Technologies) on Matrigel-coated plates. Human BJ fibroblasts (ATCC) were maintained in Eagle's minimum essential medium (EMEM) supplemented with 10% FBS and 1% Pen/Strep. V6.5 mouse embryonic stem cells (kind gift from the Sebastiano lab, Stanford) were grown in Glasgow Minimum Essential Medium (GMEM) supplemented with 10% FBS, 2 mM L-glutamine, 1% Pen/Strep, 1 mM sodium pyruvate, 1000 units/mL Leukemia Inhibitory Factor (LIF, Millipore), 1x Minimum Essential Medium NonEssential Amino Acids (MEM NEAA, Invitrogen) and 50 uM β-Mercaptoethanol. Cells were maintained on gelatin-coated dishes without feeders. All cell lines were maintained at 37° C and 5% CO_2 at recommended density and were treated and harvested at mid-log phase for

all experiments. All suspension cells were harvested using standard cell culture procedure, and adherent cells were detached using accutase (Sigma-Aldrich).

2.2 - Drug treatments

K562 cells were treated with 1 µM CDK4/6 inhibitor (PD 0332991, Pfizer) for 24 h, or with 1 µM Imatinib (Gleevec, Novartis) for 1 h, or 10 µM JNK inhibitor VIII (CAS 894804-07-0, Calbiochem) for 24 h. GM12878 cells were treated with 25 ng/mL rhTNFα (eBiosciences) for 6 h. Cell cycle analysis was performed after indicated time of treatment and a 2 h incubation in PBS, mimicking the conditions cells undergo on the microfluidics device.

2.3 - Flow cytometry analysis

In a 1.5 mL tube, cells were washed with ice cold PBS, then fixed in 1% paraformaldehyde (PFA) for 10 min followed by permeabilization using 0.5 % TritonX100 in PBS for 10 min at room temperature. Cells were stained with primary antibodies rabbit anti GATA1 (1:400, Cell Signaling, D52H6), mouse anti GATA2 (1:100, Abnova, H00002624-M01) or mouse or rabbit IgG as Isotype control in PBS containing 0.5% TritonX100, 2 mM EDTA and 0.5% BSA (Sigma) for 1 h at room temperature. After washing with staining buffer, cells were labeled with Alexa - conjugated donkey anti mouse or anti rabbit Alexa 488 or Alexa 647 antibodies (Life Technologies) at a dilution of 1:500 for 30 min at room temperature. Finally, cells were washed and analyzed using the BD FACS AriaII. For cell cycle analysis, cells were fixed and permeabilized as above, then incubated with DAPI or propidium iodide (PI) with RNAse A for 10 min before analysis.

2.4 - Immunofluorescence

The cells were washed with ice cold PBS in a 1.5 mL tube, then fixed in 1% PFA for 10 min followed by permeabilization using 0.5 % TritonX100 in PBS for 10 min at room temperature. Cells were stained with primary antibodies rabbit anti GATA1 (1:400, Cell Signaling, D52H6), mouse anti GATA2 (1:100, Abnova, H00002624-M01) or mouse or rabbit IgG as Isotype control in PBS containing 0.5% TritonX100, 2 mM EDTA and 0.5% BSA (Sigma) for 1 h at room temperature. Then cells were washed with staining buffer and incubated with secondary antibodies donkey anti rabbit Alexa 488 and goat anti mouse 594 (both 1:500) for 30 min at room temperature, then washed again. The final pellet was resuspended in 2 drops of mounting medium containing DAPI (Vectashield) and mounted on glass cover slides. Images were taken on the AxioObserver.Z1 (Zeiss) using a 40x objective and the AxioCamMR3 camera.

2.5 - PCR primers and HT barcode design

For this study we further optimized the PCR conditions and expanded the multiplexing capacity of barcoded PCR primers from 12-plex{Buenrostro:2013bc} to 8,832-plex. PCR conditions and primer sequences were optimized to produce fewer primer dimers, and therefore increase alignment percentage, without compromising PCR efficiency. We also expanded the multiplexing capacity by incorporating a dual indexing strategy. We incorporated barcode sequences from Illumina (i5 #1-8 and i7 #1-12) and designed 84 new i5 and i7 adapters. Barcode sequences were designed to have a levenshtein distance of 4 or more and sequences containing homopolymers of 4 or more were removed.

2.6 - Single-cell ATAC-seq library construction

We used the C_1 Single-Cell Auto Prep System with its Open App™ program (Fluidigm, Inc.) and an improved transposase-based library preparation strategy to perform single-cell ATAC-seq. Single cells were captured using the C_1 Single-Cell Auto Prep IFC microfluidic chips with the "ATACseq: Cell Load and Stain (1861x/1862x/1863x)" scripts (**scripts available upon request**), generated using the C_1™ Script Builder software. Prior to loading cells onto the Fluidigm IFC, cells were washed 3-5 times in C_1 DNA Seq Cell Wash Buffer (Fluidigm). Cells at a concentration of 200-250 cells/µL (K562) and 300-350 cells/µL (all others) were combined with C_1 Cell Suspension Reagent at a ratio of 3:2. 5 µL of this cell mix was loaded on to the Fluidigm IFC. Cells were stained using a green-fluorescent calcein-AM dye (LIVE/DEAD cell viability assay, Life Technologies) by diluting 2000x into the C_1 DNA-seq Cell Wash Buffer (Fluidigm) before loading onto the C_1 IFC, for K562 cells we also included a red-fluorescent ethidium homodimer-1 dye to determine viability. Single cells were captured on 96 capture sites, see supplementary table 1 for a list of experiments and IFC sizes. After cell capture, IFCs were transferred to a Leica CTR 6000 microscope for imaging.

20 µL of Tn5 transposition mix (1.5x TD buffer, 1.5 µL transposase (Nextera DNA Sample Prep Kit, Illumina), 1.5x C_1 Loading Reagent with low salt (Fluidigm), and 0.15% NP40), 20 µL of Tn5 release buffer (50 mM EDTA, 1x C_1 Loading Reagent without salt, and 10 mM Tris Buffer, pH 8), 20 µL of EDTA quenching buffer (45 mM $MgCl_2$, 1x C_1 Loading Reagent without salt, and 10 mM Tris Buffer, pH 8), and 24 µL of PCR mix (1.4 µM non-indexed custom Nextera PCR primers 1 and 2 (**Supplemental Table 1**), 1x C_1 Loading Reagent with low salt, and 1.1x NEBnext High-Fidelity PCR Master Mix) were added to the designated wells according to the "ATACseq: Sample Prep (1861x/1862x/1863x)" scripts.

ATAC-seq fragments were prepared on the Fluidigm C_1 using the script "ATACseq: Sample Prep (1861x/1862x/1863x)", running for approximately 4.5 hours. On the IFC, the Tn5 transposition reaction was carried out for 30 minutes at 37°C in the presence of the 0.1% NP40. After transposition, Tn5-DNA complexes were dissociated from chromatin by adding 50 mM EDTA for 30 min at 50°C (**Extended Data Fig. 1**). Because excess EDTA sequesters free magnesium, potentially affecting subsequent enzymatic reactions, we used 45 mM $MgCl_2$ to quench free EDTA and proceeded to perform 8 cycles of PCR using the following conditions: 72°C for 5 min; 98°C for 30s; and thermocycling at 98°C for 10s, 72°C for 30s, and 72°C for 1 min. The amplified transposed DNA was harvested in a total of 13.5 µL C_1 Harvest Reagent.

In a 96-well plate, 10 µL of harvested libraries were amplified in 50 µL PCR for an additional 14 cycles (1.25 µM custom Nextera dual-index PCR primers (**Supplementary Table 1**) in 1x NEBnext High-Fidelity PCR Master Mix) using the following PCR conditions: 72°C for 5min; 98°C for 30 s; and thermocycling at 98°C for 10 s, 72°C for 30 s, and 72°C for 1 min. The PCR products were pooled creating a final volume of ~4.8 mL. The pooled library was purified on a single MinElute PCR purification column (Qiagen) yielding libraries at an approximate concentration of ~1 µM. Libraries were quantified using qPCR prior to sequencing.

3 - Data preprocessing

3.1 - Sequencing, read trimming and alignment

All single-cell ATAC-seq libraries were sequenced using paired-end, dual-index sequencing. We used either 50x8x8x50 cycle reads from a HiSeq or 76x8x8x76 cycle reads on a NextSeq. Adapter sequences were trimmed from FASTQs using custom python scripts to enable mapping fragments with sequences containing adapters. Paired-end reads were aligned to hg19 or mm10 using BOWTIE2[1] using the parameter –X2000 allowing fragments of up to 2 kb to align. Duplicates were removed and library size was estimated using PICARD tools (http://picard.sourceforge.net). Reads were subsequently filtered for alignment quality of >Q30 and were required to be properly paired. Reads mapping to the mitochondria, unmapped contigs and chromosome Y were removed and not considered.

3.2 - Peak calling and filtering

We used MACS2[2] to call all reported ATAC-seq peaks. MACS2 was used with the following parameters (--nomodel --nolambda --keep-dup all --call-summits). Peaks were filtered using the consensus excludable ENCODE blacklist (http://hgdownload.cse.ucsc.edu/goldenPath/hg19/encodeDCC/wgEncodeMapability/). We also found a subset of high-signal regions, generally unique to ATAC-seq, which appeared to represent mitochondrial homologues. To develop a custom blacklist targeted for mitochondrial homologues, we generated synthetic 34mer reads derived from mitochondrial sequences. After mapping and peak calling of these synthetic reads we found 111 peaks in hg19 and 28 peaks in mm10. For all subsequent analysis we discarded peaks falling within these regions. Although we saw little effect on calculations of variability, in the case of K562 data, peaks were additionally filtered to exclude copy number amplifications. Using the filtered peak set, peak summits were extended +/-250 bps. The top 50,000 non-overlapping 500 bp summits, which we refer to as accessibility peaks, were used for all downstream analysis.

3.3 - Fragment analysis

As in our previous work[3], we adjusted the plus strand aligning reads by +4 and the minus strand aligning reads by -5 bp to represent the center of the transposon binding event. For calculating accessibility for each peak, we counted the number of fragments (not reads) from each cell falling into each of the 50,000 peaks. To filter cells from individual libraries, in general Calcein/EthD-1 staining gave us a qualitative measure of cell viability. We often found cases of apparently healthy cells with poor measures of accessibility (Extended Data Fig. 3) and that cell viability was sometimes difficult to assess. We therefore filtered libraries by requiring >15% of fragments falling in open chromatin (peak set defined above) and having a library size >10,000 as estimated by PICARD (Figure 1d).

3.4 - Tn5 insertion scores

We used Tn5 insertion scores as described in our previous study[3] to estimate the predicted Tn5 insertion propensity determined using the position weight matrix (PWM) of the transposase.

4 - Deviation inference

4.1 - Determining the expected signal

Analysis scripts for the deviation inference are available in the supplementary materials. In short, the deviation and variability inference algorithms were designed to

quantify whether ATAC-seq signal from a given set of accessibility peaks, defined by a specific feature of interest (such as TF binding site, replication timing domain, ChIP-seq signal etc.), varies from cell to cell more than would be expected from a set of similar accessibility peaks that do not share this common feature. First, peaks called in the aggregated data (see peak calling and filtering above) were used to determine the expected number of fragments in a given set of peaks given the number of fragments obtained for each cell. To calculate raw deviation for a given feature, we first calculated the expected signal using the following equation:

Equation #1

$$Expected\ signal = Signal_{all\ cells} * \frac{Total\ Signal_{\ cell}}{Total\ Signal_{all\ cells}}$$

where $Total\ Signal_{all\ cells}$ represents the number of fragments from all identified accessibility peaks across all cells, $Total\ Signal_{\ cell}$ represents the number of fragments from all accessibility peaks in that individual cell, and $Signal_{all\ cells}$ represents the number of fragments within the set of accessibility peaks of interest (i.e. marked by a TF motif, replication timing domain, etc.). This expected number of fragments ($Expected\ signal$), was subtracted from the observed signal in that cell ($Observed\ Signal$) to obtain a raw deviation signal (in units of fragments). We found this raw deviation signal required further correction for cell-to-cell differences in enrichment for open chromatin and Tn5 transposase sequence bias (Tn5 bias is also highly correlated with GC bias). To do this we normalized the raw deviation signal using a set of peaks with similar ATAC-seq signal intensity and with second-order corrections for transposase sequence bias. This background model is described in detail below.

4.2 - Defining the background peak set

To define our background set of similar accessibility peaks, we sampled (with replacement) peaks with similar overall accessibility scores (defined by the number of fragments in the aggregated data set), selecting the same number of peaks as was in the test set. We then determined the number of fragments expected for each cell based on these peaks ($Signal_{intensity}$). This factor provides an expected number of fragments that captures variation expected due to systematic cell-to-cell noise in the relative heights in the peaks, but not for noise generated from cell-to-cell differences in sequence-based

transposase insertion bias. To control for this potential source of technical variability, we determined a sequence-based transposase insertion bias score for each accessibility peak by calculating the mean (within the 500 bp accessibility peak) of the per-base relative Tn5 insertion probabilities (see section 2.4 Tn5 insertion scores)[3]. To determine the number of fragments observed due to variable transposase insertion bias, we calculated the number of fragments expected from peaks with similar transposition insertion bias scores as the test set ($Signal_{T-bias}$) and subtracted the number of fragments expected from this distribution of peaks given their peak intensity ($Signal_{T-bias,intensity}$) (which is also accounted for in $Signal_{intensity}$). We normalized this overall second-order transposase bias correction by ($\frac{Signal_{observed}}{Signal_{T-bias}}$) to weakly scale (on the order of 10%) the correction by the observed signal. This small correction further reduced the effects of transposase bias on the variability signal. Thus we calculate the overall expected signal accounting, non-parametrically, for bias as:

Equation #2

$$Background\ Signal = Signal_{intensity} + \left(Signal_{T-bias} - Signal_{T-bias,intensity}\right) * \frac{Signal_{observed}}{Signal_{T-bias}}$$

4.3 - Calculating normalized deviation

To calculate normalized deviation we iterated (N=30 permutations) the calculation of $Background\ Signal$. We incorporated the matched background signal that accounted for known sources of technical variation using the following equation:

Equation #3

$$deviation = \frac{Observed\ Signal - Expected\ Signal}{\sqrt{\frac{\sum_{permutations}(Background\ Signal - \quad Expected\ Signal)^2}{\#\ of\ permutations}}}$$

Henceforth and throughput the text, the reported metric of deviation represents a fold gain in signal over what is expected due to fragments from bias-compensated, matched set of peaks. To calculate variability, the metric used for cell-cell variance, we calculated the mean Background Signal over N=30 permutations per cell ($\overline{Background\ Signal}$), sum the square

of the differences for each cell, divide by the sum of the square of this normalization term, then take the square root:

Equation #4

$$variability = \sqrt{\frac{\sum_{cells}(Observed\ Signal - \quad Expected\ Signal)^2}{\sum_{cells}(\overline{Background\ Signal} - \quad Expected\ Signal)^2}}$$

Intuitively this variability metric can be thought of as the standard deviation of the observed variation in units of expected deviations of the bias-corrected background set of accessibility peaks. Thus, roughly, a variability score of 2 implies the standard deviation of the variability is twice what would be expected in the bias-correct background set.

4.4 - Error estimates and permuted definition

We estimated variability error as 1 standard deviation of the variability calculated from subsamples of $Background\ Signal$. To calculate permuted variability, a 31st permutation was calculated and then used to calculate $Background\ Signal'$ for this measure. Then permuted variability and permuted deviation were calculated as follows:

Equation #5

$$permuted\ deviation = \frac{Background\ Signal' - Expected\ Signal}{\sqrt{\frac{\sum_{permutations}(Background\ Signal - \quad Expected\ Signal)^2}{\#\ of\ permutations}}}$$

Equation #6

$$permuted\ variability = \sqrt{\frac{\sum_{cells}(Background\ Signal' - \quad Expected\ Signal)^2}{\sum_{cells}(\overline{Background\ Signal} - \quad Expected\ Signal)^2}}$$

4.5 - Quality check

To test the deviation inference algorithm, we applied the approach to features we expected to not have excess cell-to-cell variance. We partitioned peaks into deciles of their accessibility score, Tn5 bias score, and GC content, and calculated variability as described above (Extended Data Fig. 4a-f). We see near uniform variability across these features confirming that the algorithm does not capture significant accessibility deviations from these potential sources of artifactual signal.

5 - Genome annotations

5.1 - Transcription factor and histone ChIP-seq

ChIP-seq data was downloaded form the UCSC ENCODE data repository (http://hgdownload.cse.ucsc.edu/goldenPath/hg19/encodeDCC/wgEncodeAwgTfbsUniform/), analyzed using the ENCODE uniform processing pipeline which is filtered for highly reproducible ChIP peaks[4]. Histone annotations was also donwloaded from the UCSC ENCODE data repository (http://hgdownload.cse.ucsc.edu/goldenPath/hg19/encodeDCC/wgEncodeBroadHistone/) and used without further processing.

5.2 - DNA binding motifs

To map motifs to peaks we used the FIMO[5] using the parameter --thresh .00005. We mapped 208 known motifs corresponding to the JASPAR core database (2014) and motifs derived from Chen et al.[6].

5.3 - Chromatin state inferences

Chromatin states[7] were downloaded from the UCSC ENCODE data repository (http://hgdownload.cse.ucsc.edu/goldenPath/hg19/encodeDCC/wgEncodeBroadHmm/). We found ATAC-seq peaks often existed at the boundaries of the annotation. For this analysis we allowed ATAC-seq peaks to contain more than one association with a chromatin state. We also removed states 14-Repetitive/CNV and 15-Repetitive/CNV from the analysis due to concerns that variability scores were correlated with variation in karyotype between single-cells.

5.4 - Cell cycle, DNA replication

Repli-seq[8] data was downloaded from the ENCODE data repository (http://hgdownload.cse.ucsc.edu/goldenPath/hg19/encodeDCC/wgEncodeUwRepliSeq/). To assign replication timing domains to individual ATAC-seq peaks we, collected the signal values across the 6 Repli-Seq data sets (G1/S, S1, S2, S3, S4, G2/M) and assigned each

ATAC-seq peak to the data-set with the maximum signal value across the 6 replication timing data sets.

6 - Additional analysis

6.1 - Synergy score

To calculate whether TF co-association or competition created larger levels of variability we developed a synergy score reported in Figure 2e. For sets of peaks associated with factor A, factor B, and with the overlap of accessibility peaks defined as C (C = A ∩ B), we calculate the synergy score as follows: First we add the variability measured from peaks unique to A and unique to B (i.e. the disjoint set of peaks present in A and not B and B and not A) to obtain $variability_{A-C}$ and $variability_{B-C}$. We then subtract the variability measured from peaks in A and B (now containing the overlap set of peaks). To account for weak effects of peak number on variability scores, we downsampled the number of peaks associated with factor A and factor B to the number of accessibility peaks observed in A – C or B – C respectively. We re-calculated variability after downsampling 10 times and took the mean of those samples to obtain $\overline{variability}_A$ and $\overline{variability}_B$. We then calculated $SynergyScore$ using the following equation:

Equation #7

$$Synergy\ score = ((\overline{variability}_A + \overline{variability}_B) - (variability_{A-C} + variability_{B-C}))/2$$

To estimate significance values, z-scores were computed using the following equation:

Equation #8

$$z - score = \frac{SynergyScore}{standard\ dev_{permutations}(variability_{A-C} + variability_{B-C})}$$

To estimate p-values shown in Extended Data Figure 6j, we assumed that z-scores arise from a normal distribution and therefore we used a two-sided z-Test to estimate significance values.

6.2 - Biological reproducibility and drug treatments

Although we found that the deviation inference algorithm, described above, provided consistent ranking of variability across TFs, we found that the score was sensitive to sequencing depth per cell (Extended Data Fig. 4l). To quantitatively assess biological differences in variability across treatment and control, we downsampled data sets so that the treatment in question had equal numbers of reads per cell when compared to the control condition. After down sampling the input dataset, variability was calculated as described above. In analyses measuring change in variability, error bars were estimated by bootstrapping cells in treatment and control and recalculating variability and subsequently the change in variability between the cases, reported error bars represent 1 standard deviation of the calculated values from bootstrapping cells. Dotted lines shown in Fig. 3a,b, Extended Data Fig. 5a-c and Extended Data Fig. 7a,b represent a probability of detecting a change in variability >0.5 is less than 1%.

6.3 - *Cis*-correlation and chromosome conformation capture analysis

For the genome-compartmentalization analysis with scATAC-seq of GM12878s, we used all non-overlapping 500 bp accessibility peaks (n=165,067). To calculate correlation between proximal regions in the genome, we made bins of proximal accessibility peaks (N=25 peaks representing a median size of 135 kb, with step size of 10 peaks) and calculated deviation, as defined above. We note, the resolution at which previous chromosome conformation capture studies show higher-order interactions have ranged from 100 kb to 1Mb[9-11] and notably, a recent study has reported a resolution of 1 kb[12]. In this most recent work, the authors report the median contact domain to be 185 kb. We found the variability scores to be larger than expected, suggesting proximal peaks in single-cells were co-accessible. We subsequently calculated the correlation coefficient (Pearson) of every accessibility peak bin (across all cells) to all other accessibility peak bins within the corresponding chromosome. The resulting correlation values provided a correlation matrix akin to an interaction frequency measured by chromatin conformation capture assays. For comparisons to chromatin conformation capture data of GM12878 we used the interaction frequency matrix from Kalhor et al.[10] (gift from the Chen lab, USC), for comparisons to K562 we used interaction frequency matrix from Ma et al.[11] (gift from the Noble lab, UW). For subsequent comparisons, ATAC-seq data (coordinates in peaks) and chromatin conformation data (coordinates relative to restriction-enzyme cut sites) was lifted to a common coordinate system based on base-pairs.

To measure interaction compartments genome-wide, also in effort to improve our power to detect genome architecture for scATAC-seq, we employed an approach similar to previous chromatin interaction studies[9]. We used the following steps for both scATAC-seq and chromosome conformation capture data sets. First, we calculated the mean signal across the diagonal of the 2D interaction matrix (representing a 1D vector) and computed a 2D matrix background by subtracting this average 1D vector across the diagonal. To calculate a normalized 2D data matrix, we subtracted this background matrix from the observed 2D matrix, effectively nulling out the high interaction density diagonal. To calculate the *cis*-correlation score, or compartmentalization score, we calculated the correlation coefficient (Pearson) of every column of the normalized 2D matrix by every row, similar to ref. 10. The resulting matrix included negative correlation values due to anti-correlation of A/B compartments.

7 - Supplementary Discussion

7.1 - scATAC assay

Assays for understanding genome-wide chromatin dynamics from single-cells has been a missing piece in effort to understand gene regulation. In previous work[3], we developed Assay for Transposase Accessible Chromatin (ATAC-seq), a method capable of profiling chromatin accessibility from as few as 500 cells. In this work, and throughout the development of scATAC-seq, we found profiling accessibility from less than 500 cells consistently yielded poor enrichment of open chromatin (**Extended Data Fig. 2b**). From this preliminary work, we observed that the highest quality data was generally obtained when the reaction volume was scaled with the total number of cells. However, conventional methods for fluid dispensing made scaling the reaction to volumes appropriate for single cells impracticable. Using the Fluidigm Integrated Fluidics Circuit (IFC) the initial transposition reaction is carried out in 13.5 nL volume, an approximately volume-scaled reaction appropriate for maintaining the transposase number per cell ratio consistent with the previously-reported “bulk” reaction. To maximize library diversities from single-cells, we also developed transposase fragment release conditions designed to free DNA fragments into solution after the transposition reaction (see Methods). We found that the resulting libraries had improved fragment recovery per cell and maintained a high enrichment for open chromatin (**Extended Data Fig. 2b**). With single-cell ATAC-seq, cells with poor

enrichment can also be post-filtered, removing a source of noise in ATAC-seq data sets (**Fig. 1d** and **Extended Data Fig. 4d-l**).

Similar to scRNA-seq, we anticipate continued development aimed to increase recovery and throughput for our scATAC-seq method. For example, areas of potential optimization to improve recovery of ATAC-seq fragments include: 1) Tn5 concentration, 2) release conditions, potentially including SDS and other protein denaturants, 3) elimination of fragment loss due to random incorporation of transposase adapters (causing 50% of fragments to be lost due to only single species of primer being incorporated)[13]. We also find empty chambers contain approximately 100-1,000 fragments which we believe to be from cell free DNA in culture. We anticipate future improvements to the scATAC-seq protocol may also include optimized wash conditions either during cell preparation or on the IFC.

7.2 - Data analysis

Common measures of cell-to-cell variability range from low content, high-throughput immunofluorescence methods to high content, low-throughput methods such as single cell RNA-seq. These single-cell approaches generally assay either transcript levels or protein abundance, thus measuring molecular elements that often exist at high-copy number in individual cells. In contrast, regulatory genomic elements, the unit of investigation of scATAC-seq, are present at a maximum of two copies in a diploid genome. Thus the near-digital nature of accessible regulatory elements fundamentally limits assays designed to look for cellular heterogeneity at the chromatin, making interpretation of these relatively sparse, and thus noisy, signals challenging.

At first approximation, peak height within accessible regions in the genome is proportional to the fraction of cells with the active element (and is thus linked to “noise” of the element as well). This approximation has been successfully used for correlating promoter H3K27ac levels to gene expression variability[14]. For ATAC-seq, such interpretation of peak heights may be confounded given practical considerations of chromatin accessibility data. For example, an element with a longer nucleosome free region (NFR) might yield a greater number of fragments per cell than an element with a shorter NFR that is open in a larger fraction of cells. We therefore would interpret ensemble measures of accessibility as a combination of i) the fraction of cells containing the element and ii) the probability of observing a transposase insertion at that element, defined by the degree of openness if normalized for technical biases.

Despite the immense utility of ensemble measures of assaying chromatin state, little, if anything, can be done with ensemble data to infer correlated heterogeneity at sites within individual cells. We found any inference of individual genomic elements would be dominated by statistical noise, as each site was generally comprised of a majority of 0 or 1 fragment counts (**Extended Data Fig. 3**). To overcome this challenge, we reasoned that changes in the mean accessibility across trans-factor binding sites in the genome would act as a sensitive measure for regulatory variability and cellular state (**Fig. 2a**). We therefore built an analysis framework designed to measure correlated cell-cell variation within any set of genomic features, defined by an array of annotations (ie. TF motifs, ChIP-seq, replication timing, etc.).

We faced various technical and biological challenges in developing a robust measure of accessibility across single-cells. First, we found single-cells varied significantly in the total reads sequenced, stemming from both sequencing coverage and from inherent variation in library diversity (**Fig. 1d** and **Extended Data Fig. 1e-f**). Such challenges also pose a problem for measures of gene-expression within single-cells[15-17], in these gene-expression studies global changes in RNA yield are thought to arise from both technical bias and biological heterogeneity in transcript abundance. In scATAC-seq the relationship is less clear, thus for simplicity we chose to normalize by single-cell read counts, as is done for most scRNA-seq approaches[15-17], making the assumption that global heterogeneity was dominated by technical variation from sources such as fragment recovery and cell lysis efficiency. Notably, scATAC-seq data is sparse, therefore we can easily identify and remove all PCR duplicates originating from individual cells without having to use unique molecular identifiers (UMIs)[18,19], removing a common source of bias in single-cell methods. Second, we observed considerable heterogeneity in the fraction of reads within open chromatin (**Fig 1d**). This bias can substantially affect measures of accessibility, where cells with poorer enrichment have a greater fraction of reads originating from inaccessible chromatin (noise) appearing as accessible signal. Lastly, we found cells varied considerably in GC/Tn5 bias. We note that the Tn5 sequence logo is GC rich and is therefore highly correlated with GC bias, thus normalizing for either Tn5 or GC bias yields nearly identical results.

With these biases in mind, we sought to develop an analysis framework capable of inferring a bias-corrected gain or loss of accessibility within sets of peaks. To measure “deviation” and “variability” we employed a sampling approach intended to generate a matched background set of peaks for comparison across the peak list of interest (**Fig 2b**).

Practically, we found that matching peaks in both peak height and sequence bias was over-constrained, making this simple approach impractical. We therefore computed the bias for peak height and sequence content independently, removed their correlated component, and used this metric, which we called background signal (BS), as the expected deviation due to bias alone in each cell (see Methods for a detailed mathematical definition).

To more clearly demonstrate the biological meaning of the single-cell "deviation" metric we also provide comparisons with a more intuitive measure of accessibility, log2 fold change (**Extended Data Figure 4g-h**), which is not bias-corrected. Here, we see measure of deviation is highly correlated (R=0.90) to log2-fold change for GATA1, with single-cells varying up to 4-fold in peaks bound by GATA1. In contrast, sites containing the Nanog motif, which we expect little to no variability in K562 cells, again shows a strong correlation (R=0.80) with log2 fold change. However, we also find a population of outlier cells likely due to technical biases. In addition, to provide a more intuitive understanding of variability within individual peaks, we can partition cells as TF accessible or inaccessible to identify specific peaks that vary (**Extended Data Fig. 10**). In this analysis we find peaks that vary up to 30-fold across the two states. Finally, when measuring cellular variability within trans-effectors we expect the magnitudes of the measurements to reflect two prominent features of variation: 1) the fraction of cells affected, wherein a state present in 50% of cells would be measured as higher variability than a state existing in 25% or 75% of cells, 2) the size of the effect, wherein a factor with a stronger effect on a larger number of chromatin sites is expected to create larger variability measurements. For example, with chemical perturbations to the cell-cycle we find variability is correlated with the fraction of cells present at a given cell-cycle stage (**Extended Data Fig. 7**). Here, cytometry of DNA abundance, a measure of the fraction of cells at a given cell-cycle stage, can be used as an intuitive reflection of cellular variability.

We reasoned that with a robust measurement of cellular variability, i) we would see no significant variance in features known to cause artifacts in ensemble genomics assays (**Extended Data Fig. 4**), ii) we would identify changes across cell states consistent with known biology (**Fig. 3d**) and iii) we would report rational changes induced by chemical inhibitors and other perturbations (**Fig. 3a,b** and **Extended Data Fig. 7**). Throughout this work, we've taken these efforts to validate the approach. With scATAC-seq we discover trans-factors can contribute to cell-cell heterogeneity and interestingly, we find that single K562 cells vary across 7 significant components (**Fig. 2b**) suggesting a continuum of

cellular regulatory states. We also observe that peaks within single-cells co-vary in *cis*, which we show can be used to infer long-range chromatin organization (**Fig. 4**). Our overall findings suggest that ensemble measures of chromatin accessibility (peak height) are a sum of accessibilities within individual cells, which we believe is a product of fluctuations in *cis*- and trans- interactions.

Despite these findings, our analytical framework has limitations: 1) This framework is sensitive to the quality in peak set annotations. For K562s we had a large library of ChIP-seq data to draw from, greatly enhancing our ability to infer cellular state, however, such expansive data sets are not common in other cell lines or tissues. We anticipate future studies will leverage advances in TF footprinting[20,21] to better capture cellular state transitions. 2) Genomic annotations can be highly correlated, for example, GATA1 and GATA4 motifs are highly similar which can make it difficult to determine the causative trans-factor (**Extended Data Fig. 8a**). We anticipate including RNA-seq, scRNA-seq and high throughput measurements of protein concentrations might improve our ability to infer causality. However, inferring causality may prove to be more complex than simple changes in gene expression with alternative hypothesis ranging from alternative splicing to post-transcriptional modifications of TFs, areas we anticipate will make for exciting new avenues of research. 3) Measurements of co-variance in *cis* may be confounded by variability in trans and is limited to a resolution of N=25 peaks (median size is 135 kb). We anticipate there may be substantial gains in the ability to infer *cis* interactions by either masking out or leveraging the effects of trans variability to help infer long range-interactions. If paired with a higher-throughput scATAC-seq approach, improved analytical tools may enable inference of *cis* interactions at the resolution of an individual regulatory element. 4) TF deviations provide a global metric of aspects of regulatory variation, however, this framework provides little information on the behavior of individual peaks. We believe this may be directly addressed with advances in scATAC-seq throughput and new statistical methods. By assaying a large number of single cells users may begin to bin cells by their observed sub-states and measure accessibility in individual peaks. We show a proof-of-principle of this analysis in Extended Data Fig. 10.

In our previous work[3], we showed ATAC-seq data sets also include information regarding nucleosome positioning and chromatin compaction encoded in DNA fragment lengths. In ATAC-seq and scATAC-seq the generation of any sequenceable fragment suggests a substantially increased region of accessibility. Even fragments of length >300 bp

generally originate from substantially more accessible regions than the background chromatinized DNA. More specifically, large fragments often arise from two highly accessible insertion sites with an intervening less-accessible nucleosome array. In contrast, the expected ATAC-seq signal for closed chromatin is an acute absence of reads entirely. To preserve this information, we did not size select fragments for scATAC-seq. Also, in our analysis of scATAC-seq data we used all fragment lengths to increase sensitivity and because open chromatin regions also generate fragments a large length scales. However, in future work we anticipate fragment sizes may provide another data dimension to query chromatin variability within single-cells.

8 - Supplementary References

1. Ben Langmead & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat Meth* **9,** 357–359 (2012).
2. Zhang, Y. *et al.* Model-based Analysis of ChIP-Seq (MACS). *Genome Biol* **9,** R137 (2008).
3. Buenrostro, J. D., Giresi, P. G., Zaba, L. C., Chang, H. Y. & Greenleaf, W. J. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nat Meth* **10,** 1213–1218 (2013).
4. Landt, S. G. *et al.* ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia. *Genome Research* **22,** 1813–1831 (2012).
5. Grant, C. E., Bailey, T. L. & Noble, W. S. FIMO: scanning for occurrences of a given motif. *Bioinformatics* **27,** 1017–1018 (2011).
6. Chen, X. *et al.* Integration of External Signaling Pathways with the Core Transcriptional Network in Embryonic Stem Cells. *Cell* **133,** 1106–1117 (2008).
7. Ernst, J. *et al.* Mapping and analysis of chromatin state dynamics in nine human cell types. *Nature* **473,** 43–49 (2011).
8. Hansen, R. S. *et al.* Sequencing newly replicated DNA reveals widespread plasticity in human replication timing. *Proceedings of the National Academy of Sciences* **107,** 139–144 (2010).
9. Lieberman-Aiden, E. *et al.* Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science* **326,** 289–293 (2009).
10. Kalhor, R., Tjong, H., Jayathilaka, N., Alber, F. & Chen, L. Genome architectures revealed by tethered chromosome conformation capture and population-based modeling. *Nat. Biotechnol.* **30,** 90–98 (2012).
11. Ma, W. *et al.* Fine-scale chromatin interaction maps reveal the cis-regulatory landscape of human lincRNA genes. *Nat Meth* **12,** 71–78 (2015).
12. Rao, S. S. P. *et al.* A 3D Map of the Human Genome at Kilobase Resolution Reveals Principles of Chromatin Looping. *Cell* **159,** 1665–1680 (2014).
13. Adey, A. & Shendure, J. Ultra-low-input, tagmentation-based whole-genome bisulfite sequencing. *Genome Research* **22,** 1139–1143 (2012).
14. Shalek, A. K. *et al.* Single-cell RNA-seq reveals dynamic paracrine control of cellular variation. *Nature* **510,** 363–369 (2014).

15. Brennecke, P. *et al.* Accounting for technical noise in single-cell RNA-seq experiments. *Nat Meth* **10,** 1093–1095 (2013).
16. Grün, D., Kester, L. & van Oudenaarden, A. Validation of noise models for single-cell transcriptomics. *Nat Meth* **11,** 637–640 (2014).
17. Stegle, O., Teichmann, S. A. & Marioni, J. C. Computational and analytical challenges in single-cell transcriptomics. *Nature Reviews Genetics* **16,** 133–145 (2015).
18. Kivioja, T. *et al.* Counting absolute numbers of molecules using unique molecular identifiers. *Nat Meth* **9,** 72–74 (2011).
19. Islam, S. *et al.* Quantitative single-cell RNA-seq with unique molecular identifiers. *Nat Meth* **11,** 163–166 (2014).
20. Pique-Regi, R. *et al.* Accurate inference of transcription factor binding from DNA sequence and chromatin accessibility data. *Genome Research* **21,** 447–455 (2011).
21. Sherwood, R. I. *et al.* Discovery of directional and nondirectional pioneer transcription factors by modeling DNase profile magnitude and shape. *Nat. Biotechnol.* **32,** 171–178 (2014).