Deciphering variance in epigenomic regulators by k-mer factorization

Carl G. de Boer[1], Aviv Regev[1,2,3]

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

[2] Department of Biology, Massachusetts Institute of Technology, Cambridge, MA 02140, USA

[3] Howard Hughes Medical Institute, Chevy Chase, MD 20815, USA

Abstract

Variation in chromatin organization across single cells can help shed important light on the mechanisms controlling gene expression, but scale, noise, and sparsity pose significant analysis challenges. Here, we develop gkm-PCA, an approach to infer variation in transcription factor (TF) activity across samples through an unsupervised analysis of the variation in the DNA sequences associated with an epigenomic mark. gkm-PCA first represents each sample as a vector of DNA word frequencies for the DNA sequence surrounding an epigenomic mark of interest, and then decomposes the resulting matrix of *k*-mer frequencies per sample to find hidden structure in the data. This allows both unsupervised grouping of samples and identification of the TFs that distinguish groups. Applied to single cell ATAC-seq data, gkm-PCA readily distinguished cell types, treatments, batch effects, experimental artifacts, and cycling cells. The structure within the *k*-mer landscape can be further related to differentially active TFs, with each variable component reflecting a set of co-varying TFs, which are often known to physically interact. For example, in K562 cells, AP-1 TFs emerge as the central determinant of variability in chromatin accessibility through their diverse interactions with other TFs and variable mRNA expression levels. We provide a theoretical basis for why cooperative TF binding (and any associated epigenomic mark) is inherently more variable than non-cooperative binding. gkm-PCA and related approaches will be valuable for gaining a mechanistic understanding of the *trans* determinants of chromatin variability between cells, treatments, and individuals.

Introduction

Understanding how the dynamic interaction of transcription factors (TFs) and chromatin governs cell types, differentiation, and responses in a fundamental challenge. TFs recognize and bind to specific DNA sequences and, depending on the TF, can potentially affect chromatin structure and gene expression through various means including recruiting histone modifiers, chromatin remodelers, and the mediator complex. In particular, "pioneer" TFs may be able to open chromatin and, in so doing, allow other factors to bind to the now accessible DNA (1). Measurements of chromatin state, including features such as DNA accessibility, histone modifications, and TF occupancy, have shed important light on the mechanisms governing gene expression.

Epigenomic data has recently increased dramatically in scale and complexity, with studies profiling either large numbers of individuals (*e.g.* (2-7)), or using single-cell epigenomics to profile chromatin traits in individual cells. Single cell epigenomics can help discover and understand the variation in chromatin organization and gene regulation within a single cell type or in a complex cell population (8-12). In particular, single-cell ATAC-seq (scATAC-seq) allows measurement of DNA accessibility in single cells, including at high throughput (9, 10).

However, single cell epigenomics data is inherently sparse, since every locus is present at only two copies per diploid cell (9), such that ascertaining the state of an individual cell is challenging. One solution is to pool signals – either across cells (*e.g.*, of the same known type or a discovered cluster) (8) or across loci sharing a known trait (*e.g.*, binding by a

TF) (8-10). Unfortunately, rare cell states may be overlooked when common or bulk-based peaks are used as the basis for clustering or grouping (8-10), whereas clustering cells directly from single cell epigenomic data is difficult because of its sparsity (8, 10). Grouping loci by TF motifs (9) reduces this sparsity by averaging across multiple loci and, furthermore, may represent the nature of TFs interacting with chromatin. However, it requires that motifs for all relevant TFs be known *a priori*, and that these motifs faithfully represent the specificities of the TFs.

Conversely, the representation of regulatory DNA as a set of DNA words (*k*-mers) has been used extensively in the past (*e.g.*, (13-15)), and can capture TF specificities even if they are not known *a priori*. In particular, studies using chromatin profiles from bulk populations show a differential frequency of the *k*-mers associated with these marks in different cell types (16, 17). This in turn captures the differential activity of TFs and the chromatin marks they relate to, such that a cell type with a higher lever of an active TF has more of the *k*-mers it recognizes associated with the chromatin mark (**Fig. 1a** - top). This principle has been used to identify differential TF binding between samples (18). However, existing approaches are unsuitable for exploratory analysis where the identities of the samples are unknown, as may be the case for new cell subtypes or states in a population of single cells.

Here, we present gkm-PCA (gapped-*k*-mer PCA), a method that uses matrix factorization and dimensionality reduction to: (**1**) analyze variation in *k*-mer occupancy across single cells as a basis for distinguishing different cell types, states, and treatments; (**2**) identify

differentially active TFs; and (**3**) decipher TF-TF interactions. Applying gkm-PCA to scATAC-seq profiles, we show that cell-cell variation in *k*-mers associated with open chromatin is a robust and information-rich representation that can easily distinguish different cell types, drug treatments, biological artifacts, and cycling cells. Leveraging known TF specificities, we demonstrate that the individual components of our reduced-dimensionality *k*-mer space correspond to individual or groups of TFs. The TFs comprising a component are more likely to physically interact, consistent with biochemical cooperativity between TFs, which we show is expected to be especially variable. gkm-PCA thus provides a highly effective tool for exploratory data analysis for high-dimensional single cell epigenomics.

Results

gkm-PCA captures variations in *k*-mer frequency in open chromatin

Since some TFs can modify chromatin where they bind, the differential activity of TFs should be reflected in differential chromatin states at locations containing the TF's binding motif. For example (**Fig. 1a** – top), if the levels of a given active TF in a cell are too low for it to bind its motif and modify chromatin, then the chromatin modification will be not be associated with this TF's motifs. However, as the level of an active TF rises, it will bind its motif in the DNA and modify chromatin, leaving signature motifs next to the chromatin modification it elicited. Thus, by capturing motifs (represented by *k*-mers) associated with the chromatin mark, we can infer the activity of its cognate TF. In the context of chromatin accessibility **(Fig. 1a**), as the level of an active TF that opens

chromatin rises, it should bind more (**Fig. 1a** - top), opening chromatin and its binding sites in the process. Meanwhile, changes in the concentration of an active TF that cannot open chromatin has no impact on the accessibility of its binding sites (**Fig. 1a** – middle). Finally, if two TFs bind together (either because they work cooperatively, or because one potentiates binding of the other), we expect that the accessibility of their binding sites should co-vary (**Fig. 1a** – bottom). Although we may not know *a priori* what TFs are variable in a system, nor what sequences each TF recognizes, following the frequency of gapped *k*-mers (DNA words of length *k*, containing gaps) in different chromatin regions should allow us to uncover such dependencies. In particular, because a TF may recognize multiple related *k*-mers, these related *k*-mers should co-vary with each other, reflecting on the (hidden) activity of their joint, cognate TF.

To capture these dependencies in *k*-mer space we devised gkm-PCA, a procedure that combines matrix factorization with dimensionality reduction (**Fig. 1b**; **Supplementary Fig. 1**). gkm-PCA (**1**) takes as input profiles of chromatin marks or accessibility across a set of cells (or samples); and (**2**) counts, for each cell or sample, the frequencies of gapped *k*-mers (length 2-8, all possible gaps) at loci associated with a chromatin mark of interest, yielding a matrix of *k*-mer frequencies by samples. It then (**3**) decomposes this matrix of *k*-mer frequencies using PCA to identify groups of *k*-mers that co-vary across the samples and reduces the dimensionality of the data. Finally, (**4**) we can explore the relationships between cells/samples in this reduced-dimension space, and identify the *k*-mers (and associated TFs) that underlie differences between cells or samples.

gkm-PCA identifies cell types, treatments, and outliers

We applied gkm-PCA to scATAC-seq data from 1,440 single human cells, spanning drug treated and untreated cells from the chronic myelogenous leukaemia cell line K562, as well as lymphoblastoid cell lines (LCLs; GM12878), human embryonic stem cells (H1ESC), fibroblasts (BJ), erythroblasts (TF-1), and promyeloblasts (HL-60) (9). We scored *k*-mers within 50 bp of each transposon integration site (open chromatin locus; see **Methods**), decomposed the resulting *k*-mer frequency matrix using PCA, and applied t stochastic neighborhood embedding (t-SNE) to the resulting significant PCs (**Methods**) to facilitate visual inspection (**Fig. 2a**).

Cells from the different cell types readily partitioned into distinct clusters, as did cells of the same type (K562) between treatments. We also observed separation between different untreated replicates, suggesting possible batch effects with biological implications. In particular, a subset of K562 cells from one replicate formed a separate cluster (**Fig. 2a** "K562-rep3 outliers"), distinct from the other K562 cells. These outlier cells had consecutive cell indices (**Fig. 2b**), representing adjacent cells on the C1 chip, suggesting an experimental artifact.

One grouping (**Fig. 2a, "Mixed"**) was comprised of multiple distinct cell types, including some of every cell type except fibroblast (BJ) cells, and we hypothesized these may represent cycling cells sharing a common cell cycle signature. To test this, we counted the number of ATAC-seq reads in the different replication timing domains previously defined by Repli-seq in K562 cells (19) and calculated, for each cell, the ratio of reads

from (G2+S) domains to those from G1 domains (**Fig. 2c**). Cells with a high (G2+S)/G1 ATAC-read ratio either fall into the "mixed" grouping, or form a separate sub-region of a single cell type grouping, alongside the non-replicating cells of the same type (*e.g.*, HL60 cells – right side; **Fig. 2a,c**).

Principal components of accessible *k*-mer space represent differential TF activity

There are 131 significant PCs (20) in the space of accessible *k*-mers when using all cells, suggesting variation in the activities of individual or combinations of TFs between or within cell types. Specifically, we hypothesized that each PC may represent the differential activity of one or more correlated TFs or sets of TFs, capture by the relevant *k*-mers (*e.g.*, **Fig. 1a**), across cells.

To identify PC-defining *k*-mers, we examine the loadings of the *k*-mers for each significant PCs, reflecting the relative contribution of each *k*-mer to that PC (**Fig. 1b**). Next, we relate the different PCs to differential TF activity by classifying each *k*-mer into "cognate" and "non-cognate" for each TF using both the measured *in vitro* preference of each TF to individual 8-mers (PBM 8-mer Z-scores) and the generalized PWM motifs derived from it (21). Finally, we determined the enrichment or depletion of "cognate" *k*-mers among *k*-mer weights for each PC using the minimum hyper geometric statistic (**Methods**).

We applied this approach to determine differential TF activity across treated and untreated K562 cells. We performed gkm-PCA of only the K562 treated and untreated

cells in the two main K562 clusters (**Fig. 2a**; "K562-treated" + "K562-untreated"), finding 53 significant PCs, some of which located differences between treated and untreated cells (**Methods**). Both in the full initial analysis and here, the three different K562 treatments (JNK inhibition, BDR-ABL kinase inhibition (Imatinib), and CDK4/6 inhibition) yield similar partitioning of cells in accessible *k*-mer space (**Fig. 2a** and **3a**). Since PC3 and PC5 best distinguished treated from untreated cells (**Fig. 3a**), we examined the loadings of the *k*-mers for these PCs, reflecting the relative contribution of each *k*-mer to each PC (**Fig. 3b**). Whereas some *k*-mers have high loadings in both PC3 and 5 (**Fig. 3b** – top right quadrant of scatter plot), others are distinctly highly or lowly loaded in one PC but not the other (**Fig. 3b** – e.g. *k*-mers recognized by both JUND and JUNB have high loadings in PC3 and low weightings in PC5).

Relating the PCs to known specificities of human TFs, we found a large number of enriched/depleted TFs for both PC3 and PC5. In particular, the two AP-1 family TFs JUNB and JUND were enriched in PC3 and 5, respectively (**Fig. 3b**). Even though the two PWM motifs derived from the PBM data are remarkably similar for these two factors (**Fig. 3b**, bottom right), the PBM Z-scores on which these enrichments are based clearly distinguish these two PCs. Interestingly, these two motifs are enriched in open chromatin in cells treated with JNK inhibitors that prevent the activation of JUN by JNK. Since JUNB and JUND homodimers (which these PBM Z-scores represent) are not substrates for JNK (22), the decreased stability of JUN resulting from JNK inhibition may yield more JUNB and JUND homodimers, resulting in more of these homodimer binding sites in open chromatin.

PCs capture variation in TF activity across individual cells

Next, we explored TFs for variation in their inferred activity *within* a cell type, by performing gkm-PCA of only the untreated K562 cells (**Fig. 2a** – "K562-untreated"; **Methods**). Of the 27 significant PCs, 13 distinguished different replicates (**Supplementary Fig. 2**), indicating that at least some of the variability captured on these PCs represents differences between batches. We excluded these PCs from subsequent analysis, and tested the remaining 14 PCs, that showed primarily cell-cell variability (**Methods**) for enriched TFs. Overall, 40.5% (167/412) of expressed TFs with known motifs were associated with at least one gkm-PC, but this number may be inflated because many TF binding sites are so similar.

We considered some of the possible causes for the cell-cell variation in the (inferred) activity of TFs. In particular, TFs with variable activity may be more variably expressed at the RNA level, leading to cell-cell variation at the protein level, or generally lowly expressed, such that the protein level is significantly impacted by bursts of transcription. (There are, of course, other options, independent of RNA or expression levels, such as variation in upstream signaling molecules that affect the TF's activity.). To consider the first two options, we used scRNA-seq of untreated K562 cells (23) to compare the average expression levels and variability (mean corrected CV) in expression across single cells for our *k*-mer-based "variable" and "constant" TFs.

We found that the TFs that were most enriched among the PCs, and hence inferred to have the most variable activity, were expressed on average at lower levels than the least

enriched TFs (Wilcoxon rank sum test P=0.08; **Supplementary Fig. 3a**), but the two groups had a similar mean-corrected coefficient of variation (CV) (Wilcoxon rank sum test P=0.54; **Supplementary Fig. 3b**; **Methods**). Most TFs tend to have a low mean-corrected CV, with notable exceptions including the AP-1 proteins JUN, FOSL1, BATF, and ATF3 (**Supplementary Fig. 3c**).

PCs help identify TF-TF interactions

Finally, we hypothesized that different TFs that are co-enriched (or co-depleted) on the same PC could reflect dependencies or interactions between the activity of those TFs, such as cooperative binding in a complex or through one TF rendering the sites of the other accessible (**Fig. 1a** – bottom). However, because many TFs have very similar specificities and are difficult to distinguish from their cognate motifs alone, we first eliminated any motifs that closely match another more highly enriched motif (**Methods**). This was particularly important for TFs in the AP-1 family, which share very similar motifs and were often enriched together (e.g. JUN, JUNB, JUND, FOS, FOSL1, FOSB, BATF, BACH1, ATF3, SMARCC1), and are associated with five of the 13 cell-variable gkm-PCs, often in combination with other TFs.

Such analysis of individual PCs highlights putative interactions. For example, in PC13, AP-1 + SNAI3 + MAFF + SMAD3 are co-enriched (one putative interaction), whereas CTCF + NFYA are co-depleted (an opposite interaction), while PC7 represents AP-1 + IRF2/9/STAT1 (enriched) *vs*. HIC2 + other TFs (depleted) (**Supplementary Table 1**). Some of the TFs co-enriched in the same PC are known to interact with each other

physically. For instance, the AP-1 transcription factors (e.g. JUN and JUNB) are known to interact with both RUNX2 (CBFA1) (24) and SMAD3 (25) (PCs 3 and 13, respectively). In another example, interactions are also known between IRF9 and STAT1 (26) (PC7), ATF3 and JUN (27) (PC6; AP-1 motif represented by BATF motif), and the JUN factors and SPI1 (PU.1) (28, 29); (PC7; AP-1 factors represented by SMARCC1 motif). Overall, there are 2.5 times more high-confidence protein-protein interactions (30) than expected amongst TFs that are enriched together in a PC (hypergeometric test P=0.03).

Discussion

gkm-PCA provides a new approach to leverage scATAC-seq data, to partition cells by distinct epigenomic landscapes, and to understand their regulatory underpinning. It can help identify cell types and the transcriptional regulators that mediate underlying differences in chromatin. Here, we found that gkm-PCA distinguishes cell types, cycling cells, and experimental artifacts, and discovered a large number of significant PCs in all datasets analyzed, each appearing to represent one or more TFs.

One possible explanation for the variation in inferred TF activity across single cells is variation in the expression of the TF between the cells, as has been previously shown by scRNA-seq, RNA-FISH, and single cell protein staining (e.g. (31-33); reviewed in (34)). However, we found that TFs associated with cell-cell epigenomic variability across untreated K562 cells are relatively lowly expressed in all cells, but not particularly variable across cells, as reflected by scRNA-seq. One possible explanation is that

variation would be more apparent post-transcriptionally, such as in protein translation, modification, or stability, either because of direct regulation of these steps or because of separation of time scales. Consistent with this possibility, low mRNA expression levels generally result in more variable (noisier) protein levels (35) since transcription or decay of a single mRNA results in greater fold differences in low-abundance genes. An alternative explanation is that a TF would show variable binding dependent on a variable co-factor, while itself not being variable (e.g. **Fig. 1a** - bottom).

The primary axes of variation in the K562 scATAC-seq data, as reflected by the gkm-PCs, appear to represent the combined actions of multiple TFs, often known to interact physically. This may reflect cooperative binding by these TFs. Cooperative binding mediated by physical interaction between TFs (**Supplementary Fig. 4**) or by mutual competition with nucleosomes (36) results in a steeper binding curve, such that small changes in concentration around the critical point result in larger changes in occupancy than in a non-cooperative setting. Thus, cell-cell variability in TF concentration around this point will result in higher occupancy/accessibility variability than would be expected in the non-cooperative case.

Cooperativity may also provide some insight into the prevalence of AP-1 factors in our analysis, whose binding sites were enriched in many gkm-PCs for both treatment-associated and cell-variable PCs. AP-1 TFs are bZIP TFs and can form a large number of heterodimers with other bZIP TFs (29), some of whose motifs were also found to be enriched on the same gkm-PCs as the AP-1 factors. The strong enrichment of AP-1

motifs in variable *k*-mer axes associated with scATAC-seq indicates that AP-1 factors may themselves be associated with mediating chromatin accessibility. Indeed, it has been suggested previously that AP-1 factors have pioneer activity (37).

A remaining challenge – present whenever motifs are used to infer TF binding – is the definitive identification of causal TFs when many TFs have similar motifs and the specificities of many TFs remains unknown (21). One advantage of a *k*-mer-based approach is that much of the analysis can be done without ever knowing the identities or specificities of the TFs. In this way, our knowledge deficits regarding TF binding specificities are shifted from the analysis to the interpretation stage, knowing that the specificities themselves can be captured in *k*-mer space. As we learn more about how TFs function, our interpretation of the *k*-mer space will improve.

As cell numbers grow, it is likely that additional insights can be gleaned, allowing us to detect less variable TFs and TF-TF combinations. We anticipate that gkm-PCA will also be useful in the study of other chromatin profiles collected across single cells (*e.g.*, scChIP-seq (8)). It can also help understand variation in chromatin organization in the analysis of many bulk samples, for example, those collected across individuals in a population (*e.g.*, (2-7)). Although other *k*-mer based methods have been applied to study of variation in *cis* (18), we anticipate that the unsupervised approach of gkm-PCA will be useful in dissecting variation in *trans*. With epigenomic data of ever increasing complexity, tools and approaches like those we described will continue to provide insight into the regulation of chromatin.

Methods

Data processing

A summary of the data processing steps and tools used is included in **Supplementary Fig. 1**, and a bash script with exact commands, computational tools, and useful R functions are available on GitHub (https://github.com/Carldeboer/GKM-PCA).

Data was obtained from the Gene Expression Omnibus, accession GSE65360. Samples were demultiplexed, and reads trimmed for Nextera adaptors and mapped to the human genome (hg19) using Bowtie2 (38) using paired reads (-X 2000), as described previously (9). Regions of interest were defined as windows of 50 bp to either side of the 5' end of mapped reads, representing the integration sites of the Tn5 transposase, merging overlapping regions (which removes duplicate reads). DNA sequences were then extracted from these loci using twoBitToFa (39) and scanned for *k*-mer content using AMUSED (https://github.com/Carldeboer/AMUSED), considering both DNA strands, to yield a vector of *k*-mer frequencies for each cell that was used in subsequent analyses, including all gapped *k*-mers from length 2 to 8. We stopped at a length of k=8 because for k>8 *k*-mer frequencies become very sparse when analyzing as few loci per cell as are present in scATAC-seq data, although larger k may be more suitable to analysis of bulk samples. Cells with fewer than 3,162 ($10^{3.5}$) distinct Tn5 integration loci were excluded from subsequent analyses to remove dead cells and cells with poor data quality.

The individual cells' *k*-mer frequency vectors were merged and scaled so that each *k*-mer had mean 0 and a standard deviation (SD) of 1, and this matrix was decomposed into its

principal components. For all analyses, PCA was done with the prcomp R function and the number of significant PCs was estimated using the permutationPA function for the jackstraw R package (20), tSNE was done using the tsne R package using the default parameters and including only the significant PCs. Because the frequencies of *k*-mers of varying G+C-content are so correlated to G+C content itself, the first PC often has a significant G+C-content component and should be analysed carefully (*e.g.*, GG tends to occur more frequently with higher G+C-content, and so the two will be correlated and both will be anticorrelated with A+T-rich *k*-mers).

Scoring cells for cell cycle signatures

Using the ENCODE Repli-seq data for K562 cells (19), the genome was divided into replication domains using a percent signal cutoff of 25%, where any region with a signal greater than this cutoff was considered a domain for the respective stage of the cell cycle. ATAC-seq reads were then counted within each domain to yield a matrix of ATAC-seq read counts for each domain in each cell. This matrix was scaled by the total number of reads per cell, yielding a matrix of proportions of reads per domain per cell, and the ratio of (G2+S1+S2+S3+S4)/G1 (termed (G2+S)/G1 above) was used to distinguish cycling cells.

Identifying PCs that distinguish treated from untreated K562 cells

Every cell was "scored" by its position as it is projected onto the respective PC axis. The area under the ROC curve (AUROC) statistic and rank sum P-value, representing how well the projected cell positions divide the cells into treated and untreated cells, were

calculated, and the PCs with the AUROC furthest from 0.5 (*i.e.* those for which treated cells are either enriched or depleted by the PC) were considered those that segregated treated from untreated best.

Identifying TF-specific PCs

Ungapped 8-mer protein binding microarray Z-scores and position weight matrices (PWMs) for all human TFs (inferred or directly determined) were downloaded from CIS-BP (21). For PWMs, gapped *k*-mer scores were derived by finding the maximum log-odds score for that *k*-mer in the PWM. These scores were then converted into Z-scores by centering them about the median and scaling them to the median absolute deviation, taking a Z-score of >3 as "bound" and leaving others as "unbound" *k*-mers. For PBM Z-scores, Z-scores between experiments for the same TF were combined using Stouffer's method and those *k*-mers with a Z-score above 3 were considered "bound", with others "unbound".

With this set of "bound" and "unbound" *k*-mers for each TF, the enrichment of each TF in each PC axis was calculated using the minimum hypergeometric test (40). Briefly, the bound and unbound *k*-mers were ranked by their PC weights and, moving in increasing rank order, hypergeometric P-values were calculated representing the enrichment of *k*-mers bound by that TF amongst the top N most highly (lowly) weighted *k*-mers. Exact P-values (considering the dependence between tests) were not calculated and instead multiple hypothesis testing correction using Bonferroni's method was done as if the tests were independent, yielding a more conservative P-value (to minimize the number of non-

specific TF enrichments). For PBM Z-scores, only the top 3,000 *k*-mers were considered, while for PWM scores it was the top 15,000 *k*-mers (because these also included gapped *k*-mers and was approximately the same percent of all *k*-mers). Only TFs expressed in K562s were considered (41).

Because many TFs share similar *k*-mer binding profiles and the number of *k*-mers considered for PWM motifs was so high, these appeared to have a high false positive rate and so we set the threshold for significance much lower for PWM motifs ($P<10^{-112}$) than for 8-mer Z-scores ($P<10^{-2}$). ($\log_{10}$(P-values) are "inflated" with PWMs as a result of common shared submotifs and a very large number of gapped k-mers; we chose these cutoffs based on the "elbow" of the log-P-value distributions, which are similar at these values.) To eliminate redundant motifs and select only the most enriched of each group of related motifs, the most enriched (or depleted) motif was retained and any redundant motifs (*k*-mer Pearson R > 0.5) were eliminated until all TFs were either eliminated due to redundancy or selected to represent the PC.

Comparison to K562 single-cell RNA-seq

A matrix of single cell count data was downloaded from GEO (GSE90063) for wild type K562 cells (23) and a negative binomial distribution was fit to the gene-wise mean and variance, representing a theoretical minimum variance dependent on the mean, and this was used to calculate the theoretical minimum log coefficient of variation (CV). We then subtracted the theoretical minimum CV from the observed log CV per gene to get the excess CV over that expected from its dependence on the mean ("mean-corrected CV").

We then compared the distributions of the mean-corrected CV and expression mean for TFs that had a significant enrichment among the cell-variable PCs and those that did not, using the Wilcoxon rank sum test. Cell-variable PCs excluded any PCs that significantly distinguished any replicate from the other two (Bonferroni-corrected Wilcoxon rank sum test $P < 0.1$), and also excluded PC1 because of the association with G+C content.

TF cooperativity occupancy

As described previously (42), a TF's (x) fractional occupancy of a single binding site (O_x) depends on its concentration ($[x]$) and the dissociation constant (Kd_x) of its binding site in the following relationship, which represents 1 minus the probability the binding site will *not* be bound:

$$O_x = 1 - \frac{1}{1 + \frac{[x]}{Kd_x}}$$

If TF x can also bind with a partner y, occupancy of x depends on x binding in isolation, as before, but also binding with y as a xy heterodimer, depending on the concentration $[xy]$ and the Kd_{xy} of the heterodimer. At equilibrium, $[xy] = [x][y]Ka_{xy}$, where Ka_{xy} is the association constant of x and y. Thus, for x binding to a single binding site with or without cooperative binding of y, we have:

$$O_x^{coop} = 1 - \left(\frac{1}{1 + \frac{[x](1 - Ka_{xy}[y])]}{Kd_x}}\right)\left(\frac{1}{1 + \frac{[x]Ka_{xy}[y]]}{Kd_{xy}}}\right)$$

For simplicity, we can assume that [y] is constant since the same logic holds if x and y are interchanged and for arbitrary [y]. Thus, $Ka_{xy}[y]$ is a constant corresponding to the fraction of x that is in xy form. Assuming $Kd_{xy} < Kd_x$ (since xy has both x and y binding DNA, and so is expected to bind more tightly), as [x] changes, this cooperative occupancy is always at least as steep as without cooperativity at concentrations yielding intermediate occupancy, regardless of choice of parameters, resulting in saturation of binding over a shorter range of [x] with cooperative binding. Intuitively, this is because increasing [x] increases cooperative and non-cooperative binding equally when $Kd_{xy} = Kd_x$, but when $Kd_{xy} < Kd_x$ cooperative binding increases more rapidly until saturation. **Supplementary Fig. 4** was made assuming 1% of x is in xy form, and Kd_{xy} is 100x lower than Kd_x.

Acknowledgements

We thank Jason D. Buenrostro, Christine S. Cheng, Marcin Tabaka, Nir Friedman, and Karthik Shekhar for helpful discussions and careful review of the manuscript, Atray Dixit and Karthik Shekhar for help with the K562 scRNA-seq data, and Leslie Gaffney for help with figures. C.G.D. is supported by a Canadian Institutes for Health Research Fellowship. AR is an HHMI Investigator. Work was supported by a CEGS grant from NHGRI and by HHMI.

References

1. Magnani L, Eeckhoute J, & Lupien M (2011) Pioneer factors: directing transcriptional regulators within the chromatin environment. *Trends in genetics : TIG* 27(11):465-474.

2. Sui WG, *et al.* (2014) ChIP-seq analysis of histone H3K9 trimethylation in peripheral blood mononuclear cells of membranous nephropathy patients. *Brazilian journal of medical and biological research = Revista brasileira de pesquisas medicas e biologicas / Sociedade Brasileira de Biofisica ... [et al.]* 47(1):42-49.
3. Sui W, *et al.* (2013) Genome-wide analysis of histone H3 lysine9 trimethylation by ChIP-seq in peripheral blood mononuclear cells of uremia patients. *Hemodialysis international. International Symposium on Home Hemodialysis* 17(4):493-501.
4. Rendeiro AF, *et al.* (2016) Chromatin accessibility maps of chronic lymphocytic leukaemia identify subtype-specific epigenome signatures and transcription regulatory networks. *Nature communications* 7:11938.
5. Cheng CS, *et al.* (2016) Genetic determinants of chromatin accessibility and gene regulation in T cell activation across human individuals. *bioRxiv*.
6. Sun W, *et al.* (2016) Histone Acetylome-wide Association Study of Autism Spectrum Disorder. *Cell* 167(5):1385-1397 e1311.
7. Chen L, *et al.* (2016) Genetic Drivers of Epigenetic and Transcriptional Variation in Human Immune Cells. *Cell* 167(5):1398-1414 e1324.
8. Rotem A, *et al.* (2015) Single-cell ChIP-seq reveals cell subpopulations defined by chromatin state. *Nature biotechnology* 33(11):1165-1172.
9. Buenrostro JD, *et al.* (2015) Single-cell chromatin accessibility reveals principles of regulatory variation. *Nature* 523(7561):486-490.
10. Cusanovich DA, *et al.* (2015) Multiplex single cell profiling of chromatin accessibility by combinatorial cellular indexing. *Science* 348(6237):910-914.
11. Jin W, *et al.* (2015) Genome-wide detection of DNase I hypersensitive sites in single cells and FFPE tissue samples. *Nature* 528(7580):142-146.
12. Clark SJ, Lee HJ, Smallwood SA, Kelsey G, & Reik W (2016) Single-cell epigenomics: powerful new methods for understanding gene regulation and cell identity. *Genome biology* 17:72.
13. Zhang MQ (1998) Identification of human gene core promoters in silico. *Genome research* 8(3):319-326.
14. Jensen LJ & Knudsen S (2000) Automatic discovery of regulatory patterns in promoter regions based on whole cell expression data and functional annotation. *Bioinformatics* 16(4):326-333.
15. Blanchette M & Tompa M (2002) Discovery of regulatory elements by a computational method for phylogenetic footprinting. *Genome research* 12(5):739-748.
16. Setty M & Leslie CS (2015) SeqGL Identifies Context-Dependent Binding Signals in Genome-Wide Regulatory Element Maps. *PLoS computational biology* 11(5):e1004271.

17. Ghandi M, Lee D, Mohammad-Noori M, & Beer MA (2014) Enhanced regulatory sequence prediction using gapped k-mer features. *PLoS computational biology* 10(7):e1003711.
18. Lee D*, et al.* (2015) A method to predict the impact of regulatory variants from DNA sequence. *Nature genetics* 47(8):955-961.
19. Consortium EP (2012) An integrated encyclopedia of DNA elements in the human genome. *Nature* 489(7414):57-74.
20. Chung NC & Storey JD (2015) Statistical significance of variables driving systematic variation in high-dimensional data. *Bioinformatics* 31(4):545-554.
21. Weirauch MT*, et al.* (2014) Determination and inference of eukaryotic transcription factor sequence specificity. *Cell* 158(6):1431-1443.
22. Karin M, Liu Z, & Zandi E (1997) AP-1 function and regulation. *Current opinion in cell biology* 9(2):240-246.
23. Dixit A*, et al.* (2016) Perturb-Seq: Dissecting Molecular Circuits with Scalable Single-Cell RNA Profiling of Pooled Genetic Screens. *Cell* 167(7):1853-1866 e1817.
24. D'Alonzo RC, Selvamurugan N, Karsenty G, & Partridge NC (2002) Physical interaction of the activator protein-1 factors c-Fos and c-Jun with Cbfa1 for collagenase-3 promoter activation. *The Journal of biological chemistry* 277(1):816-822.
25. Liberati NT*, et al.* (1999) Smads bind directly to the Jun family of AP-1 transcription factors. *Proceedings of the National Academy of Sciences of the United States of America* 96(9):4844-4849.
26. Horvath CM, Stark GR, Kerr IM, & Darnell JE, Jr. (1996) Interactions between STAT and non-STAT proteins in the interferon-stimulated gene factor 3 transcription complex. *Molecular and cellular biology* 16(12):6957-6964.
27. Hai T & Curran T (1991) Cross-family dimerization of transcription factors Fos/Jun and ATF/CREB alters DNA binding specificity. *Proceedings of the National Academy of Sciences of the United States of America* 88(9):3720-3724.
28. Bassuk AG & Leiden JM (1995) A direct physical association between ETS and AP-1 transcription factors in normal human T cells. *Immunity* 3(2):223-237.
29. Chinenov Y & Kerppola TK (2001) Close encounters of many kinds: Fos-Jun interactions that mediate transcription regulatory specificity. *Oncogene* 20(19):2438-2452.
30. Rolland T*, et al.* (2014) A proteome-scale map of the human interactome network. *Cell* 159(5):1212-1226.
31. Shalek AK*, et al.* (2013) Single-cell transcriptomics reveals bimodality in expression and splicing in immune cells. *Nature* 498(7453):236-240.
32. Trapnell C*, et al.* (2014) The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nature biotechnology* 32(4):381-386.

33. Warren L, Bryder D, Weissman IL, & Quake SR (2006) Transcription factor profiling in individual hematopoietic progenitors by digital RT-PCR. *Proceedings of the National Academy of Sciences of the United States of America* 103(47):17807-17812.
34. Tanay A & Regev A (2017) Scaling single-cell genomics from phenomenology to mechanism. *Nature* 541(7637):331-338.
35. Taniguchi Y*, et al.* (2010) Quantifying E. coli proteome and transcriptome with single-molecule sensitivity in single cells. *Science* 329(5991):533-538.
36. Mirny LA (2010) Nucleosome-mediated cooperativity between transcription factors. *Proceedings of the National Academy of Sciences of the United States of America* 107(52):22534-22539.
37. Sheffield NC*, et al.* (2013) Patterns of regulatory activity across diverse human cell types predict tissue identity, transcription factor binding, and long-range interactions. *Genome research* 23(5):777-788.
38. Langmead B & Salzberg SL (2012) Fast gapped-read alignment with Bowtie 2. *Nature methods* 9(4):357-359.
39. Karolchik D*, et al.* (2004) The UCSC Table Browser data retrieval tool. *Nucleic acids research* 32(Database issue):D493-496.
40. Zilberstein CB-Z, Eskin E, & Yakhini Z (2004) Using expression data to discover RNA and DNA regulatory sequence motifs. *Proceedings of the First Annual RECOMB Satellite Workshop on Regulatory Genomics.*:65–78.
41. Pollen AA*, et al.* (2014) Low-coverage single-cell mRNA sequencing reveals cellular heterogeneity and activated signaling pathways in developing cerebral cortex. *Nature biotechnology* 32(10):1053-1058.
42. Granek JA & Clarke ND (2005) Explicit equilibrium modeling of transcription-factor binding and gene regulation. *Genome biology* 6(10):R87.

Supplementary Tables

Supplementary Table 1: Summary of TFs associated with the different untreated K562 cell-variable gkm-PCs. TFs are listed in decreasing order of enrichment

1 significance, with TFs filtered for redundancy between motifs as described in the

2 **Methods**.

PC	TFs enriched in highly weighted *k*-mers	TFs enriched in lowly weighted *k*-mers
PC3	RUNX2, RREB1, TERF2, SMARCC1, ZNF524, KLF3, SREBF1, TBX15, TWIST1	TBP, FOXD2, E2F2, POU4F1, NFATC1, IRF9, HOXD13, MEF2C, STAT5B, ZNF384, TEAD3, CDC5L, FOXP1, YY1, E2F4, LCOR, SOX12, FUBP1, SPI1, PRDM4, BBX, MLL, HES2, E2F1, HHEX, SP6, CIC
PC5		
PC6	BATF, NFE2L2, TGIF1, ATF3	
PC7	IRF2, SPI1, SMARCC1, ELF1, SPIB, IRF9, STAT1	HIC2, ZNF740, KLF1, ZNF143, MZF1, HOXB4
PC13	NFYA, CTCF	JUNB, SNAI3, MAFF, SMAD3
PC14	MZF1, ZNF740, GATA1, CREB1	
PC26	JUN, RUNX2	
PC27	ESRRA	

3

4

5

Figure Legends

Figure 1: gkm-PCA. A) The relation between the differential activity of TFs that open chromatin and the numbers of their cognate motifs associated with open chromatin. Shown is a cartoon example of the impact of TFs (circles) on chromatin accessibility when the TF's concentration is low (left) or high (right), for different scenarios of TFs that can (top and bottom rows) or cannot (middle row) open chromatin. If the TF can open chromatin either alone (top) or cooperatively (bottom), a change in the concentration or activity of TFs will affect the number of accessible binding sites in the cell (colored bars). If a TF that has no effect on accessibility (middle) there will be no relationship between accessible motifs (bars) and the TF's concentration. **B**) gkm-PCA method. From left: genomic sequences associated with open chromatin or another feature of interest are used as input (left), and the frequency of each *k*-mer in open chromatin/feature (row) is counted in each sample (column) (middle), the resulting *k*-mer frequency matrix is then decomposed by PCA (right) into the *k*-mers contributing to each PC (left matrix) and the projection of the samples into the new (PC) space (right matrix).

Figure 2: gkm-PCA identifies cell types, drug treatments, cycling cells, and experimental artifacts in scATAC-seq data. A) Identification of cell types. t-SNE two dimensional projection of the 131 significant PCs for all cells. Cells are colored by pre-annotated type (legend) and major cell type clusters are encircled. **B**) Detection of outliers. Shown are the cell indices (position on C1 chip) for cells from K562-replicate 3, with outlier K562 cells (as in A) marked in black. The outlier cells have consecutive indices suggesting a shared location on the chip. White: cells filtered out prior to analysis.

1 **C**) Cell cycle phases. t-SNE projection as in A, but with color indicating cell cycle stage
2 as determined by the ATAC reads falling within replication domains, showing that the
3 "mixed" population from A are comprised primarily of replicating cells.

4 **Figure 3: PCs represent TF variation. A**) Partitioning cells by treatment. Shown is a
5 projection of treated (shades of blue) and untreated (shades of pink) K562 cells onto PC 3
6 and 5 from a gkm-PCA of only K562 cells. **B**) Identification of TFs associated with
7 specific PCs. Scatter plot shows the PC weights for each 8-mer (dot) for PC 3 (x axis)
8 and PC5 (y axis). Colored dots: *k*-mers recognized by JUNB (red), JUND (blue), and
9 both (green), with consensus JUN 7-mer shown as a pink star, as defined using PBM 8-
10 mer Z-scores (21); the legend (bottom right) shows PWMs derived from the same PBM
11 8-mer Z-scores. Side graphs show the Log2 fold enrichment of JUNB- and JUND-bound
12 *k*-mers amongst lowly-weighted PC *k*-mer weights for PC 3 (bottom) and PC 5 (right).

13 **Supplementary Figure 1: gkm-PCA computational pipeline.** A bash pipeline and
14 other computational resources are available on GitHub
15 (https://github.com/Carldeboer/GKM-PCA). Tools/functions used are indicated in
16 brackets.

17 **Supplementary Figure 2: PCs that distinguish replicates.** Shown are the Bonferroni-
18 corrected p-values (y axis) and AUROC values (x axis) for how well each PC separates
19 each untreated K562 replicate from the other two replicates. Colors indicate the replicate

being compared to the other two. Red horizontal line: Cutoff (0.1) below which PCs were considered to separate batches.

Supplementary Figure 3: The TFs enriched in gkm-PCs have lower expression. A,B) ÇDF of the mean (population) expression (A, x axis) or mean-corrected CV (B, x axis; **Methods**) for the most (blue) and least (pink) significant TFs enriched in the PCs from a gkm-PCA of untreated K562 cells. **C**) The relationship between the mean expression (x axis) and CV (y axis) for all genes in WT K562 data (dots). Names of TFs with the highest mean-corrected CV are labeled and AP-1 factors are bolded. Pink, blue: TFs with least and most significant PC enrichment.

Supplementary Figure 4: Cooperativity between TFs results in steeper binding curves. The predicted fractional TF occupancy (y axis) for a given concentration of the TF (x axis), when the concentration of the cooperatively-interacting TF is constant. The two binding curves are aligned at 50% occupancy to emphasize the differences in the slopes. Modeling was done as described in **Methods**.

Figure 1

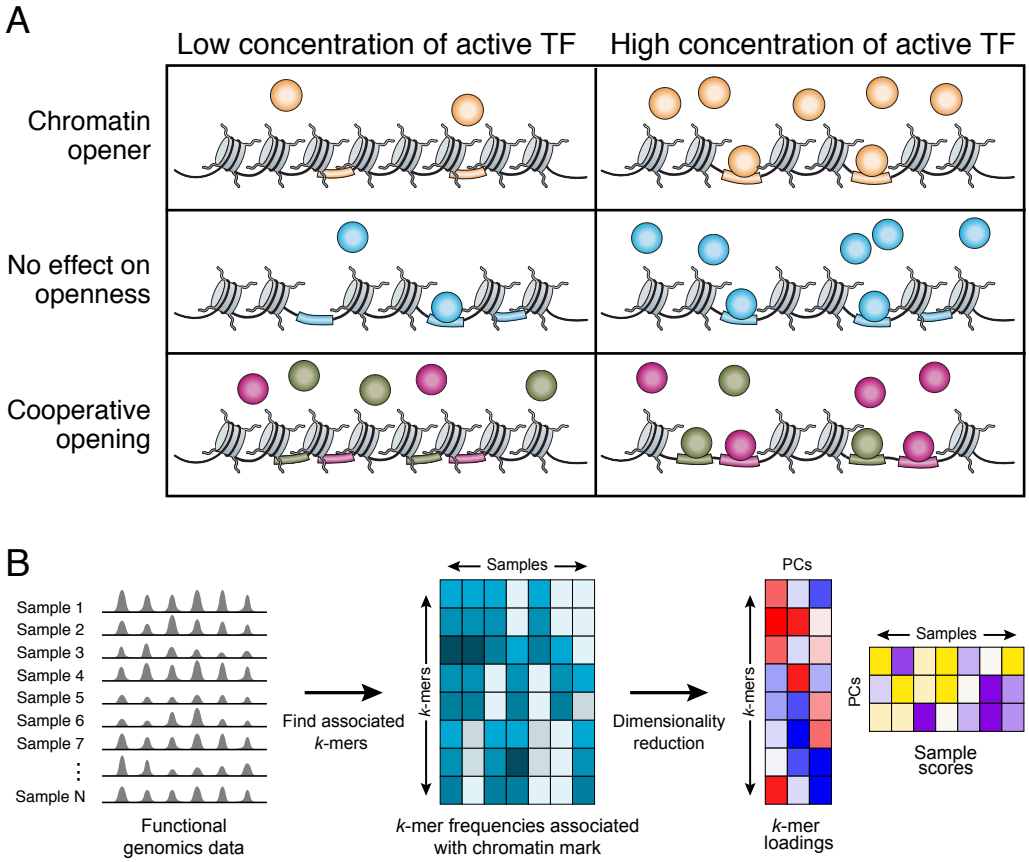

A
Low concentration of active TF
High concentration of active TF
Chromatin opener
No effect on openness
Cooperative opening
B
Sample 1
Sample 2
Sample 3
Sample 4
Sample 5
Sample 6
Sample 7
Sample N
Functional genomics data
Find associated k-mers
Samples
k-mers
k-mer frequencies associated with chromatin mark
Dimensionality reduction
PCs
k-mers
k-mer loadings
Samples
PCs
Sample scores

Figure 2

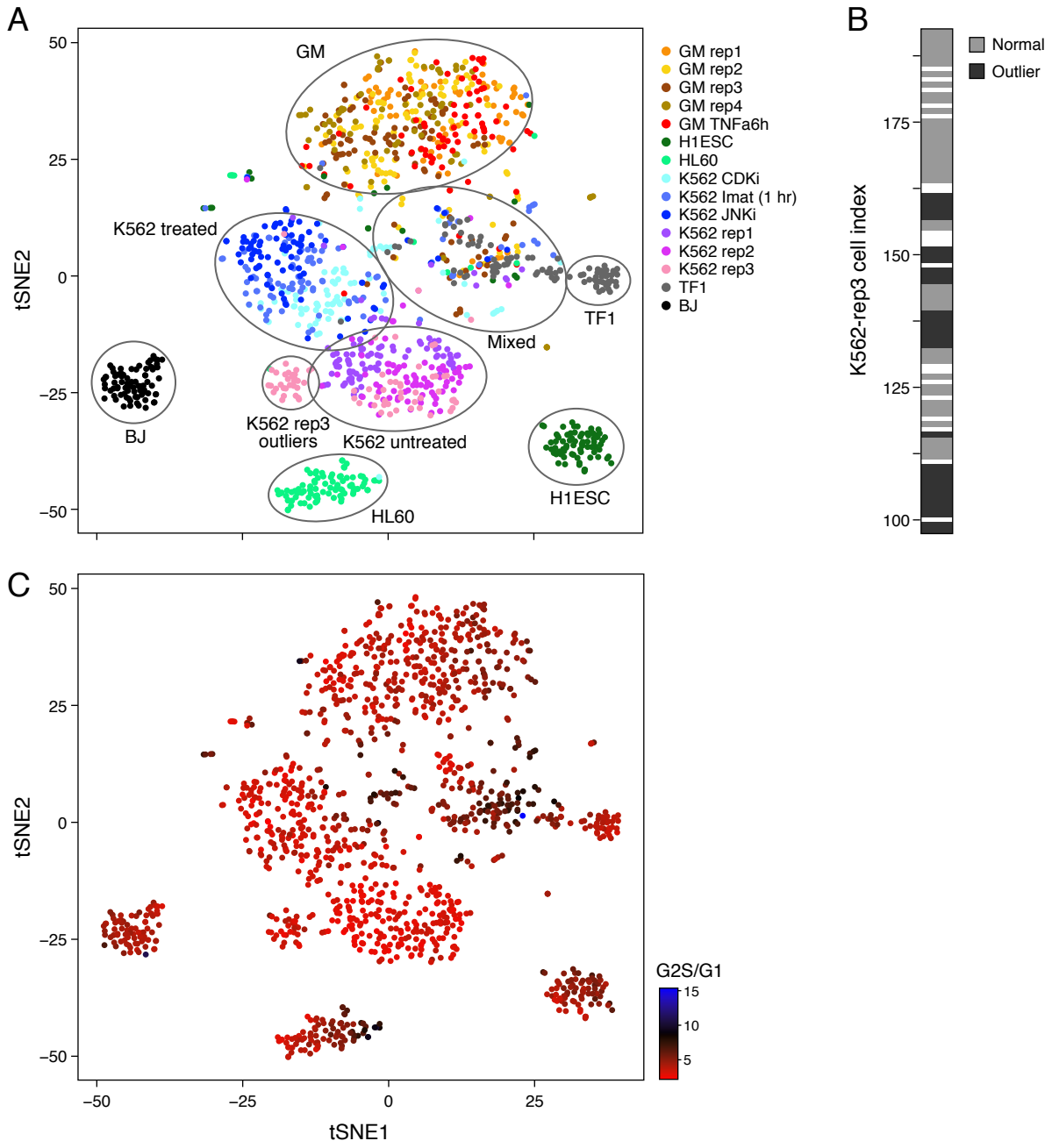
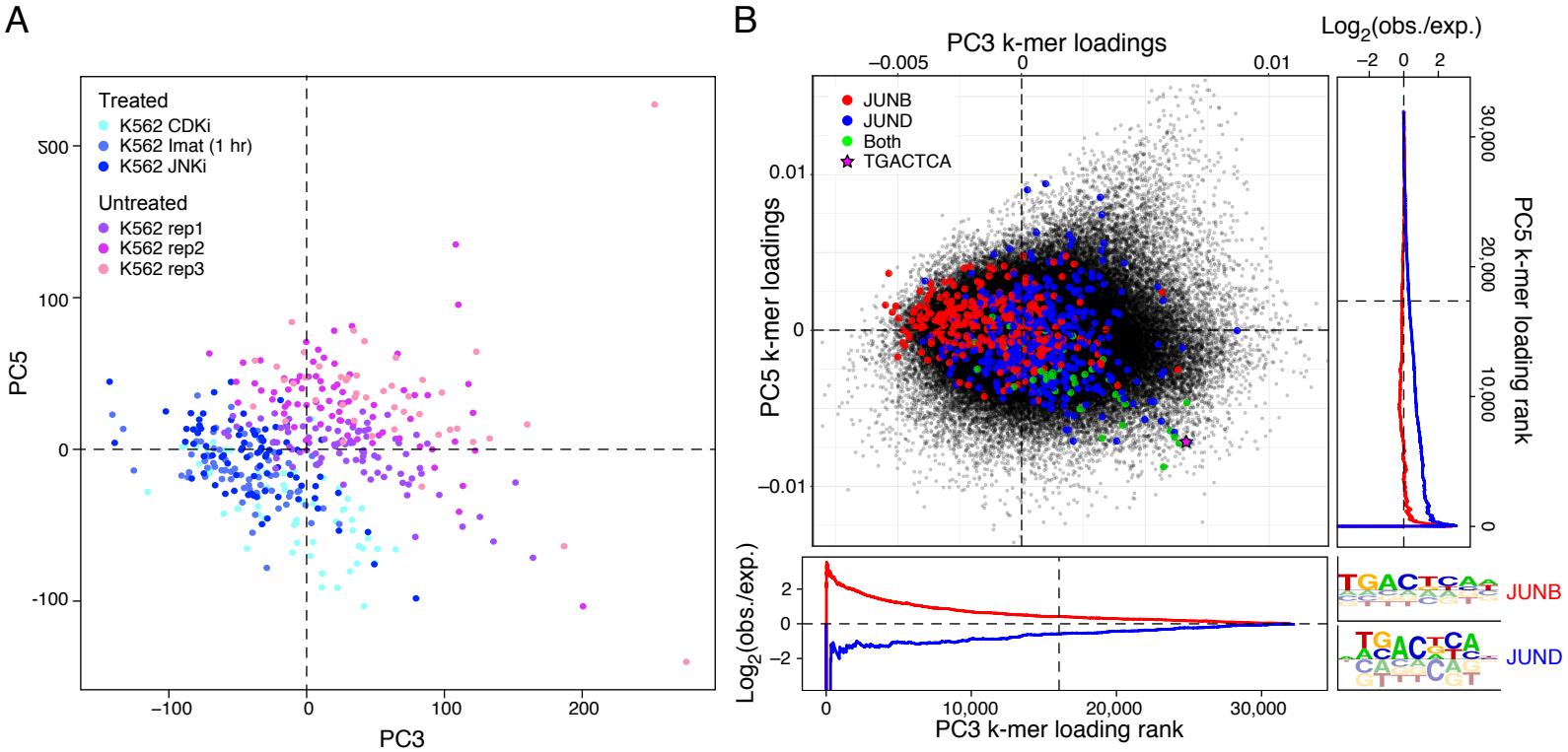
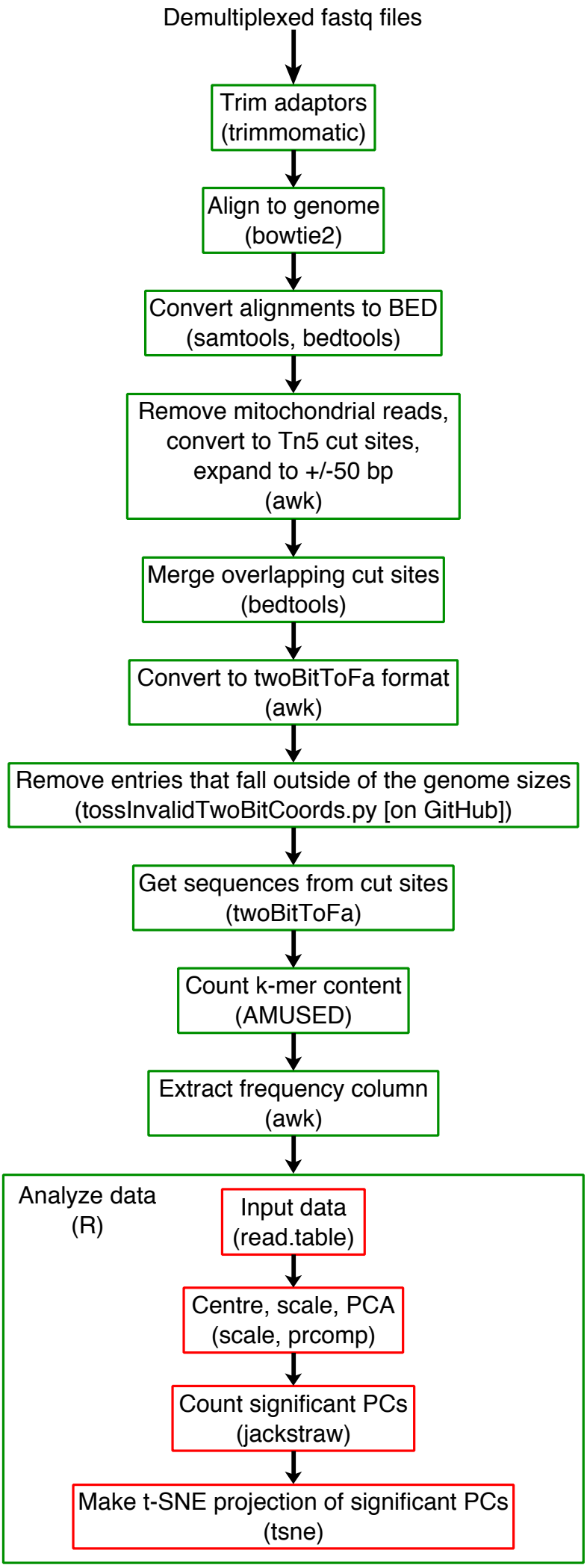
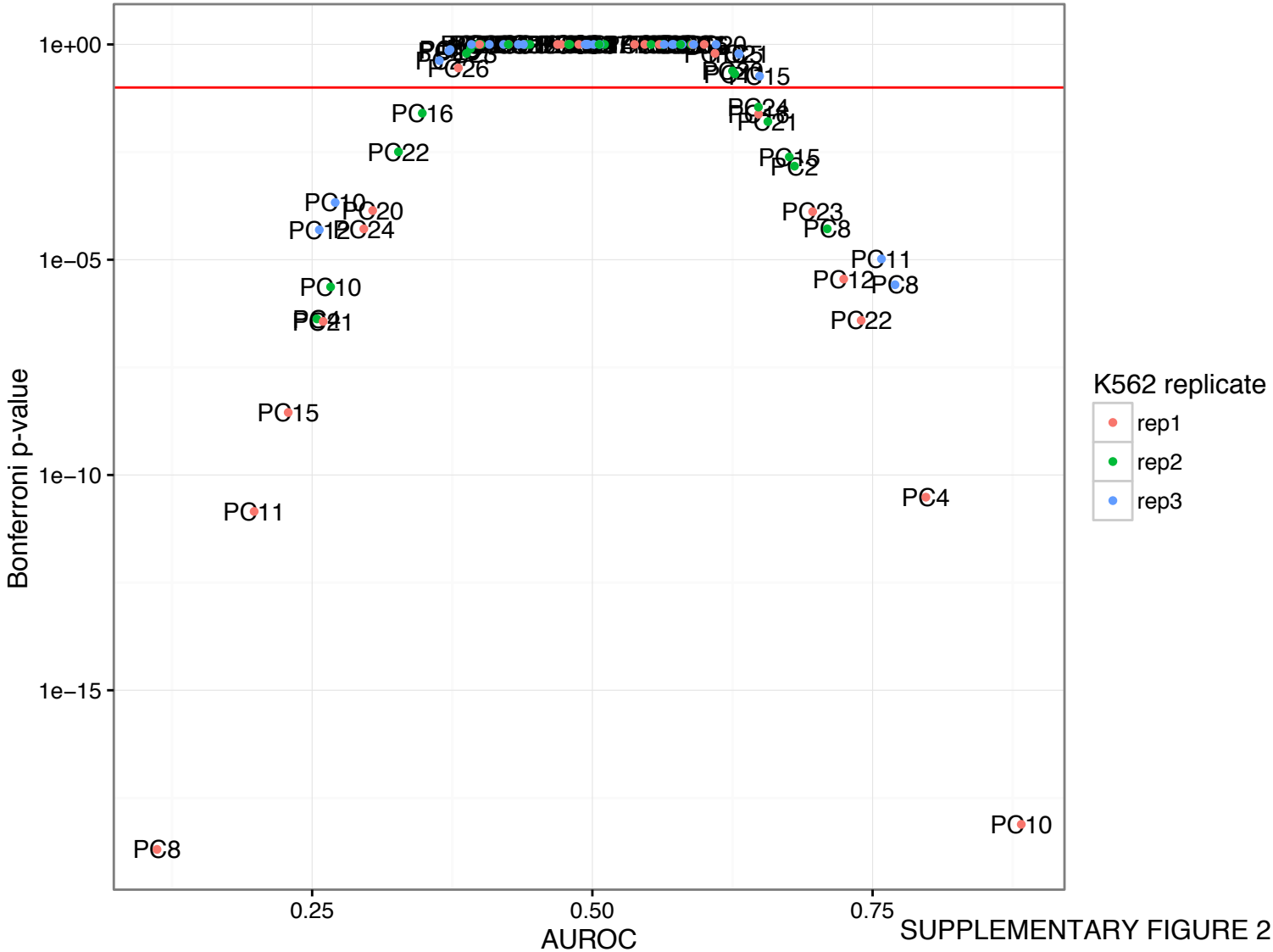

A
GM
K562 treated
Mixed
TF1
BJ
K562 rep3
outliers
K562 untreated
HL60
H1ESC
tSNE2
GM rep1
GM rep2
GM rep3
GM rep4
GM TNFa6h
H1ESC
HL60
K562 CDKi
K562 Imat (1 hr)
K562 JNKi
K562 rep1
K562 rep2
K562 rep3
TF1
BJ
B
Normal
Outlier
K562-rep3 cell index
C
G2S/G1
tSNE1

Figure 3

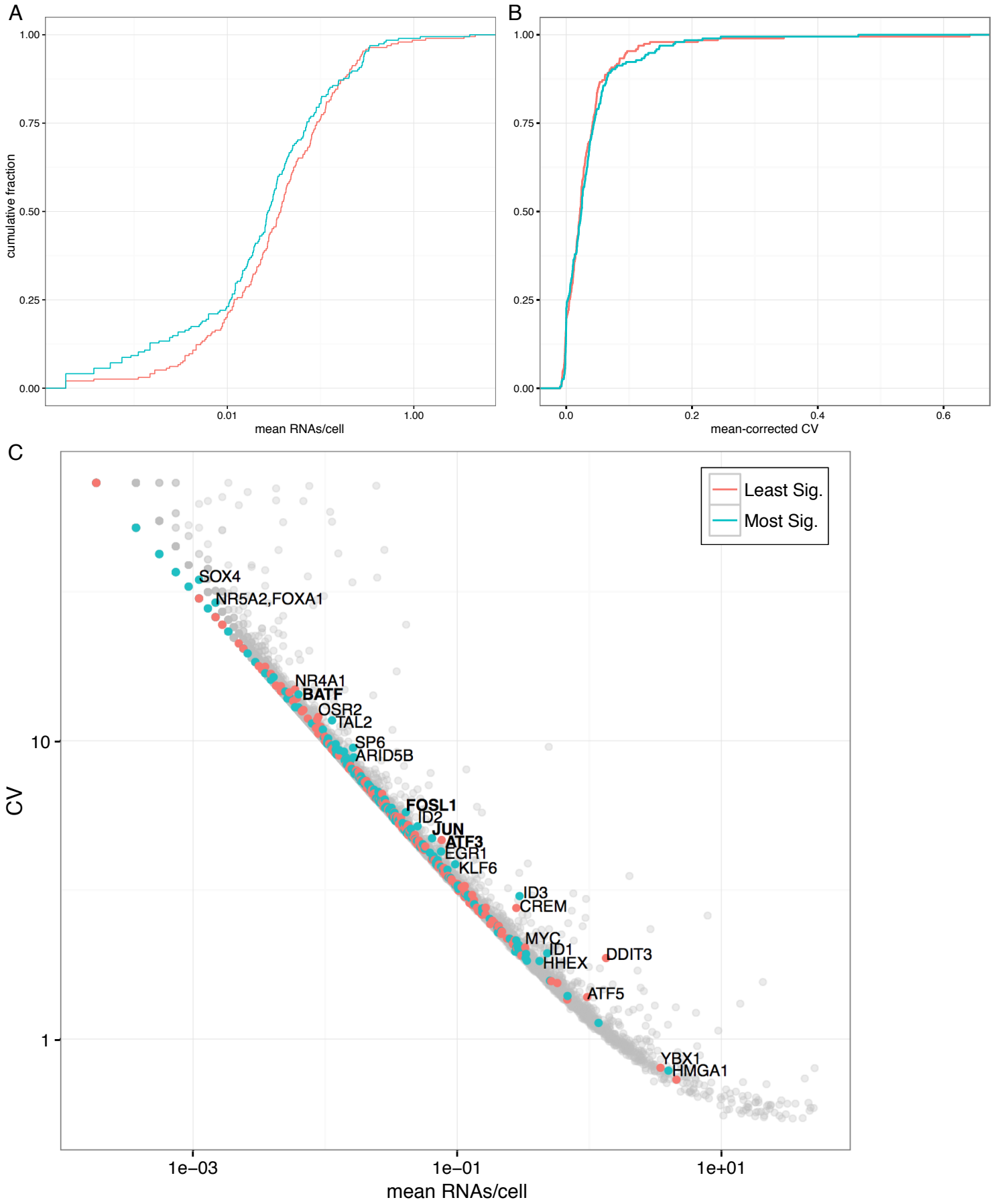

A
Treated
K562 CDKi
K562 Imat (1 hr)
K562 JNKi
Untreated
K562 rep1
K562 rep2
K562 rep3
PC5
PC3
B
PC3 k-mer loadings
Log2(obs./exp.)
JUNB
JUND
Both
TGACTCA
PC5 k-mer loadings
PC5 k-mer loading rank
PC3 k-mer loading rank
JUNB
JUND



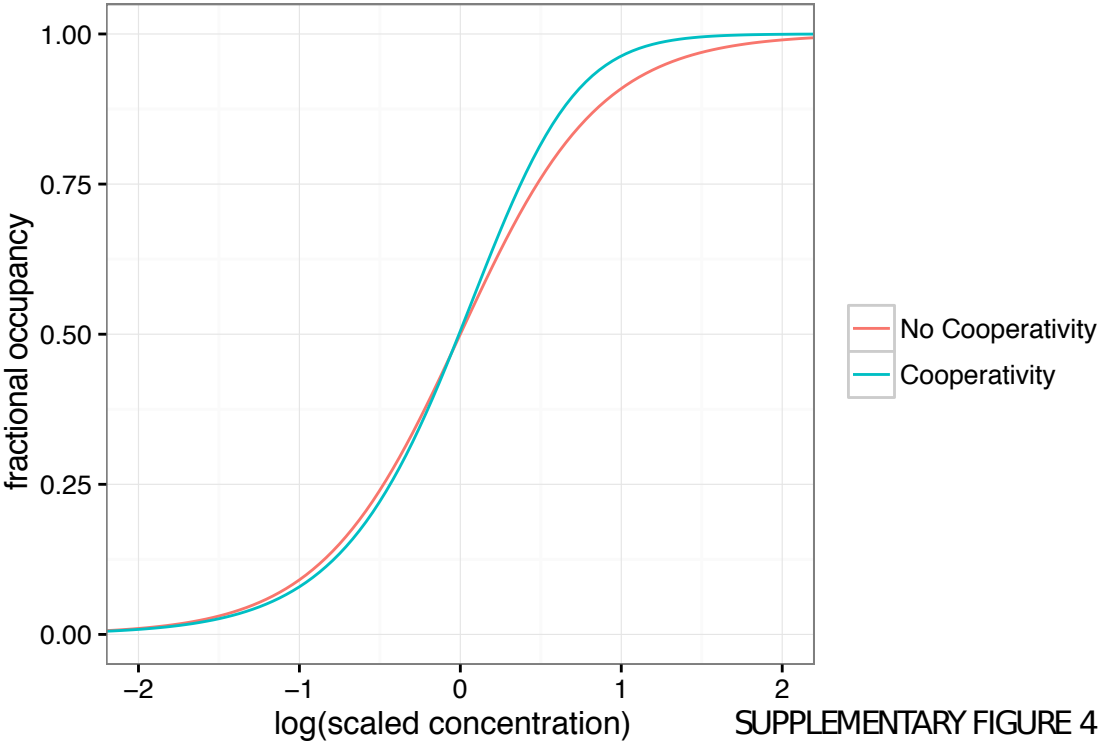
SUPPLEMENTARY FIGURE 1



SUPPLEMENTARY FIGURE 2



SUPPLEMENTARY FIGURE 3



SUPPLEMENTARY FIGURE 4