MAGICACT: A tool for predicting transcription factors and cofactors driving gene lists using ENCODE data.

Roopra, A

Dept. of Neuroscience
5507 WIMR
1111 Highland Avenue,
University of Wisconsin-Madison
Madison,
WI 53705, USA
Phone: 1 (608) 265 9072
asroopra@wisc.edu

ABSTRACT.
Transcriptomic profiling is an immensely powerful hypothesis generating tool. However, accurately predicting the transcription factors (TFs) and cofactors that drive transcriptomic differences between samples represents a challenge and current approaches are limited by high false discovery rates. This is due to the use of TF binding sequence motifs that, due their small size, are found randomly throughout the genome, and do not allow discovery of cofactors. A second limitation is that even the most advanced approaches that use ChIPseq tracks hosted at sites such as the Encyclopedia Of DNA Elements (ENCODE) assign TFs and cofactors to genes via a binary designation of 'target', or 'non-target' that ignores the intricacies of the biology behind transcriptional regulation.
ENCODE archives ChIPseq tracks of 169 TFs and cofactors assayed in 91 cell lines. The algorithm presented herein, **M**ann whitney **A**nalysis of **G**ene **C**ohorts for **A**ssociated **C**ofactors and **T**ranscription factors (MAGICACT), uses ENCODE ChIPseq data to look for statistical enrichment of TFs and cofactors in gene bodies and flanking regions in gene lists. When compared to 2 commonly used web resources, o-Possum and Enrichr, MAGICACT was able to more accurately predict TFs and cofactors that drive gene changes in 3 settings: 1) A cell line expressing or lacking single TF, 2) Breast tumors divided along PAM50 designations and 3) Whole brain samples from WT mice or mice lacking a single TF in a particular neuronal subtype. In summary, MAGICACT is a standalone application that runs on OSX and Windows machines that produces meaningful predictions of which TFs and cofactors are enriched in a gene list.

Introduction.

Key to the control of gene expression is the level of transcript in the cell. Transcription factors (TFs) are DNA binding proteins that recognize specific sequence elements (motifs) associated with gene promoters and enhancers. TFs recruit cofactors that do not themselves bind DNA but are brought to promoters via TFs to either enhance or repress gene expression. TFs and cofactors (here on termed 'Factors') are thus key regulators of transcript levels.

It is now routine to obtain the levels of every gene transcript in the genome i.e. whole transcriptome data. The datasets contain tens of thousands of expression values per sample and the number of samples can be in the thousands such as breast cancer transcriptomes archived at The Cancer Genome Atlas (TCGA)(1).
When comparing transcriptomes from two or more conditions such as normal to cancerous tissue, thousands of mRNA levels can change. We posit that in many cases, the majority of those changes are driven by alterations in the function of a few Factors that coordinate programmatic gene changes on a genome wide scale. Identifying those driving Factors is a fundamental problem and therapeutic opportunity, yet current tools are extremely poor at making such predictions.

Current algorithms fall into two categories. The first looks for motifs that are over-represented in the upstream regions of a list of genes with altered expression compared to the gene population. This approach often fails because with very few exceptions, transcription factor motifs are small (around 6bp) and often have redundant bases. A random 6bp sequence occurs approximately 750,000 times in the human genome. Thus, searches will yield a high false positive rate; most predicted sites will not be functional. Moreover, in many genes, TFs act in concert such that the presence of a single motif element is uninformative. Motif searching also precludes any prediction of cofactors because cofactors do not bind DNA directly. This is especially problematic for translational research where uncovering drugable targets is a major goal, and cofactors are often enzymes that can make excellent drug targets. The second approach relies on assigning genes to Factors (TFs and cofactors) by looking at ChIPSeq tracks such as those archived at ENCODE(1). Genes are ranked by ChIP signal and then an arbitrary number of 'top' genes are taken as targets. Though this allows for assignment of cofactors, it is problematic because it fails to take into account the biology of Factor regulation of genes. The main concern is that many Factors have highly non-gaussian ChIP signal distributions; the distributions can have large numbers of sites with very low but detectable signals and skewed, long tails of increasing ChIP signal. This thwarts unbiased attempts to define genes as 'targets' or 'non-targets'(2). A new method is required that takes into account the biology of transcriptional regulation in a statistically rigorous manner.

We have devised a novel algorithm that meshes whole transcriptome data with whole genome factor binding (ChIP Seq) data archived at ENCODE. The algorithm is termed , ***M****ann whitney* ***A****nalysis of* ***G****ene* ***C****ohorts for* ***A****ssociated* ***C****ofactors and* ***T****ranscription factors* (MAGICACT). MAGICACT circumvents the principal confounds of current methods to identify Factors, namely: 1) unacceptably high false positive rates due to the use of TF motif searches 2) inability to identify cofactors due the absence of any binding site motifs 3) arbitrary assignment of Factors to genes based on hard cutoffs of ChIPSeq signals. MAGICACT accepts an input list of genes and compares ChIP signals for the list with the population of ChIP signals for each Factor in ENCODE. We compared MAGICACT to a commonly used TF prediction alorithm, oPOSSUM-3(3) using whole transcriptome data from a REST knockdown cell line, TCGA breast tumors and mouse brain tissue. In all cases, MAGICACT predicted more biologically meaningful TFs and associated cofactors compared to oPossum.

Methods

MCF7 RNA profiling.

MCF7 cells harboring a REST shRNA knockdown and controls were generated as described(4). RNA was extracted in triplicate from shControl and shREST cells using TRIZOL, checked on an Agilent NanoChip and subjected to microarray hybridization onto Nimblegen arrays (2006-08-03_HG18_60mer_expr) at the University of Wisconsin-Madison Biotechnology Center. Spot intensities were RMA normalized(5). Probes were considered present in the experiment if they were 'present' in all 3 samples of either controls or knockdown cells. If a probe was present in only one condition, the values in the other condition were accepted regardless of the present/absent call. Probes were then collapsed to a single gene value per sample by taking the median value of all probes per gene per sample. Fold changes between control and REST knock-down cells were calculated along with Benjamini-Hochberg corrected Student t-test p values. 'Up-regulated' and 'down-regulated' gene lists for MAGICACT analysis were generated by taking any genes changed more than 3 Standard Deviations from the mean fold change and having a corrected p value<0.05.

TCGA profiling.

Breast cancer RNAseq data (data_RNA_Seq_v2_expression_median.txt) was downloaded from cBioPortal (latest datafreeze as of 6July2017). PAM50(6) designations were used to parse tumors into Luminal-A or Basal subtypes (supplemental table 9 – list of samples and PAM50 designation). A gene with an RPKM ≥1 in at least 50% of either basal or Luminal-A samples was considered for further analysis. Fold changes, statistics and gene lists were then generated as above for REST knock-down cells.

Sams et al profiling.

RNAseq data (GSE84175) was downloaded from GEO. A gene with an RPKM ≥1 in all controls or all CTCF KO cells was considered for further analysis and fold changes, statistics and gene lists were then generated as above for REST knock-down cells.

MAGICACT.

MAGICACT aims to determine whether genes in a query list are associated with higher ChIP values than expected by chance for a given transcription factor or cofactor based on ChIPseq tracks archived at ENCODE. To do this, we assigned a ChIP value for each factor in ENCODE at every gene. We then generated an algorithm to test the hypothesis that a list of query genes are enriched for high ChIP signals.

Generation of the MAGIC Matrix.

We defined a single ChIP signal for each factor in ENCODE at every gene as follows:
All genes in the human genome (hg19) were assigned a gene domain that was defined as 5Kb either side of the gene body (5'UTR – 5Kb to 3'UTR + 5Kb). NarrowPeak files for all transcription factors and cofactors were downloaded from http://genome.ucsc.edu/cgi-bin/hgFileUi?db=hg19&g=wgEncodeAwgTfbsUniform. These files provide called peaks of signal enrichment based on pooled, normalized (interpreted) data. Custom Python scripts were then used to extract the highest ChIPseq peak value (signalValue) for a given factor seen in any cell line within every gene domain. Thus each factor and gene were assigned a single signalValue – the highest value – observed in ENCODE. The highest value was chosen to 1) maximize the likelihood that this site will be bound in cell types not present in ENCODE and 2) Work by others suggest that the strongest sites may be the most functional (2). The resulting array (factor columns vs gene rows) was normalized such that the sum of all signalValues for any factor (columns) was arbitrarily set at 100,000. We term the final array the MAGIC Matrix (Supplemental File 1).

Defining Enriched Factors in a Gene list.

Two lists are entered into MAGICACT. The first is a master list containing all genes of relevance to the experiment (for example, all genes deemed to be present or expressed in an experiment). The second list contains genes of interest – the query list (e.g. genes induced above a threshold in one condtion over control).

Query lists are subsets of the master list and both lists are filtered by MAGICACT for genes present in the MAGIC matrix.

For each Factor in the Matrix, MAGICACT orders the N genes in the accepted master list by ChIP signal (lowest to highest) obtained from MAGIC Matrix. It then generates a normalized cumulative distribution function (CDF) such that:

$$F_M(c) = \frac{1}{N}\sum_{i=1}^{N} I_{[0,c]}(M_i)$$

Where $F_M(c)$ is the fraction of genes in the Master list (N genes) with ChIP signal less than c, and M are the ordered ChIP signals. $I_{[0,c]}(M_i)$ is an indicator function that is equal to 1 if $M_i \leq c$ else equal to 0. An emperical distribution function, $F_Q(c)$ is generated by similarly ordering the X query genes Q, where $Q \subset M$, and $I_{[0,c]}(Q_i)=1$ if $Q_i \leq c$ else $I_{[0,c]}(Q_i)=0$.:

$$F_Q(c) = \frac{1}{X}\sum_{i=1}^{X} I_{[0,c]}(Q_i)$$

It then screens for Factors where the query distribution is right-shifted compared to the master CDF. This is accomplished by calculating the difference between the 2 distributions and defining the supremum (D_s) and infimum D_I of the difference as:

$$D_S = sup_c\ (F_M(c) - F_Q(c))$$

$$D_I = inf_c\ (F_M(c) - F_Q(c))$$

If the cumulative of query samples is left shifted compared to the population ($|D_S| < |D_I|$), the factor is triaged and not considered further. If the query cumulative is right shifted compared to the population cumulative i.e. $|D_S| > |D_I|$, MAGICACT performs a 1-tailed Mann Whitney Wilcoxen (MWW) test between the master and query list to test the null hypothesis that the query list is drawn randomly from the master.

A Score for each Factor is calculated to incorporate the Benjamini-Hochberg corrected MWW p value, as well as a measure of how the highest ChIP values in the query list compare to the highest values in the population. Thus, for each Factor, the mean of the top n chip signals in the query list is compared to the mean of the top n ChIPs in the population to produce a ratio where:

$$n = 0.05X$$

and

$$r = \frac{\mu_q}{\mu_m}$$

Where μ_q is the mean for the top n signals for Q and μ_m is the mean of the top n signals for M.

A Score (S) is assigned to each Factor thus:

$$S = -\log\,(P_{corr}) \times r$$

Where P_{corr} is the Benjamini-Hochberg corrected MWW 1-tailed p value. Factors are then sorted by Score.

Implementation of MAGICACT.
A tab or comma delimited text file is requested by MAGICACT (lists file). The first column contains a list of all genes of relevance to the experiment (master list). This could be a list of all genes expressed in the system for example, or all genes with a detectable signal. Any number of other columns are then added that contain query lists. For example, a query list may be all genes that go up under some criterion and another may be all genes that go down. The first row is the header and must have unique names for each column.

Using the algorithm above, MAGICACT analyzes each query list and generate a series of output files and sub-directories in the directory containing the lists file. An 'Accepted_Lists.txt' file is generated which is the original lists file filtered for genes in the MAGIC Matrix. A Platform_Matrix.txt file is the MAGIC Matrix filtered for genes in the the master list. A series of sub-directories are generated named after each query list in the lists file. Each directory contains sub-directories and files for the stand-alone analysis of that query list. A query_list_summary.pdf contains a bar graph of Factors and Scores with $P_{corr} < 0.05$ (e.g. Fig. 1b). Query_list_Summary.xls is a spreadsheet containing statistical information for all non-triaged factors. Data reported in Query_list_Summary.xls are:

Critical ChIP:	ChIP value at d_{sup} (This value is used to determine driver genes in the list)
Obs Tail Mean:	Average of the 5th percentile ChIP values (n values) in the query list.
Exp Tail Mean:	Average of the top n ChIP values in the population.
Tail Enrichment:	Ratio of the Obs and Exp Tail Means (r)
Raw P:	MWW P value
Corrected P (FDR):	Benjamini Hochberg corrected p value (P_{corr})
Score:	Score ($-\log(P_{corr})$x r)

All Factors associated with $P_{corr} < 0.05$ are highlighted in bold red.

Query_list_Drivers.gmx is a tab delimited file in the GMX format utilized by GSEA(7). Columns contain all target genes of a Factor whose ChIP signal is greater than the Critical ChIP i.e. the ChIP at which there is the maximal difference between the population and query cumulative. These are considered 'driver genes' for that Factor. The GMX format will allow this file to be used directly in any subsequent GSEA analysis. The first column contains all genes in the analysis. The second column contains genes in the input list. All subsequent columns contain driver genes for the Factor at the top of the column.

A sub-directory called 'CDFs' contains graphical displays of the analysis for all factors with $P_{corr} < 0.05$, named 'rank'_'factor'.pdf (e.g. 1_NRSF.pdf; rank = 1, factor = NRSF) (Fig. 1c) where ranking is determined by Score. Two cumulative functions are displayed: the black curve is the fractional cumulative of all genes in the master list against the ChIP values, red is the same for query genes. A blue vertical line denotes the ChIP value at d_{sup}. Red ticks represent each gene in the query list and black ticks are all genes in the population. Red ticks with circles ('lollipops') are those n query genes in the 5th percentile of ChIP values. Black lollipops are genes in the population with the n highest ChIP values.

A second sub-directory called 'Auxiliary_Files' is populated with data behind the summary files. The 'query_list_raw_results.txt' file contains the same columns as the Query_list_Summary.xls file but has raw data for all factors including those that were triaged and not considered for further statistical analysis. It also contains the MWW U statistic for each factor. U statistics with a negative sign denote U values for triaged factors; the negative sign is used by the algorithm for triage sorting.

Query_list_Sub_Matrix.txt is the MAGIC Matrix filtered for genes in the query list.

Triaged_Factors.txt is a list of factors that were not considered (triaged) because $|d_{inf}| > |d_{sup}|$

Triaged_Genes.txt contains all genes in the query that were not in MAGIC Matrix and therefore eliminated from analysis.

A sub-directory named 'driverData' contains comma separated text files for each Factor with $P_{corr} < 0.05$. The file contains the list of driver genes for that Factor and associated MAGIC Matrix ChIP value.

Obtaining MAGICACT

MAGICACT can be downloaded from https://go.wisc.edu/magicact.

Results

The purpose of MAGICACT is to identify the transcription factors and cofactors that are responsible for the major patterns of gene expression changes in a disease situation or other biologic pertubation. To test MAGICACT, we asked whether MAGICACT could predict which transcription factor was knocked down in a cell line. We have shown that REST is a tumor suppressor whose loss drives tumor growth of MCF7 cells in mouse xenograft model of breast cancer. We have previously published on the transcriptome of MCF7 cells infected with lentivirus expressing either a scrambled shRNA or shRNA targeting REST(4,8,9). This simple system consists of a clonal cell line lacking a single factor. RNA was subjected to hybridization to Nimblegen arrays, and all genes were assigned a Fold Change (FC=log2(shREST/shScramble)) and those with FC>3 Standard Deviations from the mean FC and FDR<0.05 were designated as the 'Upregulated' gene list upon loss of REST (119 genes) or 'Downregulated' (35 genes not including REST itself). These 154 genes were able to distinguish shScramble from shREST samples in unsupervised heirarchical clustering (Fig.1A). Using 15,446 expressed and 119 upregulated genes (supplemental table 7), MAGICACT predicted REST as the highest scoring Factor associated with upregulated genes in shREST cells (raw $p=9.2x10^{-49}$, FDR = $2.0x10^{-47}$, Score = 45.7). MAGICACT also classified the REST corepressors SIN3A, CoREST, HDAC2, and CtBP2 as significant at FDR<0.05.(Fig.1B and supplemental table 1) (see discussion). Figure 1C shows the highly divergent cumulatives of

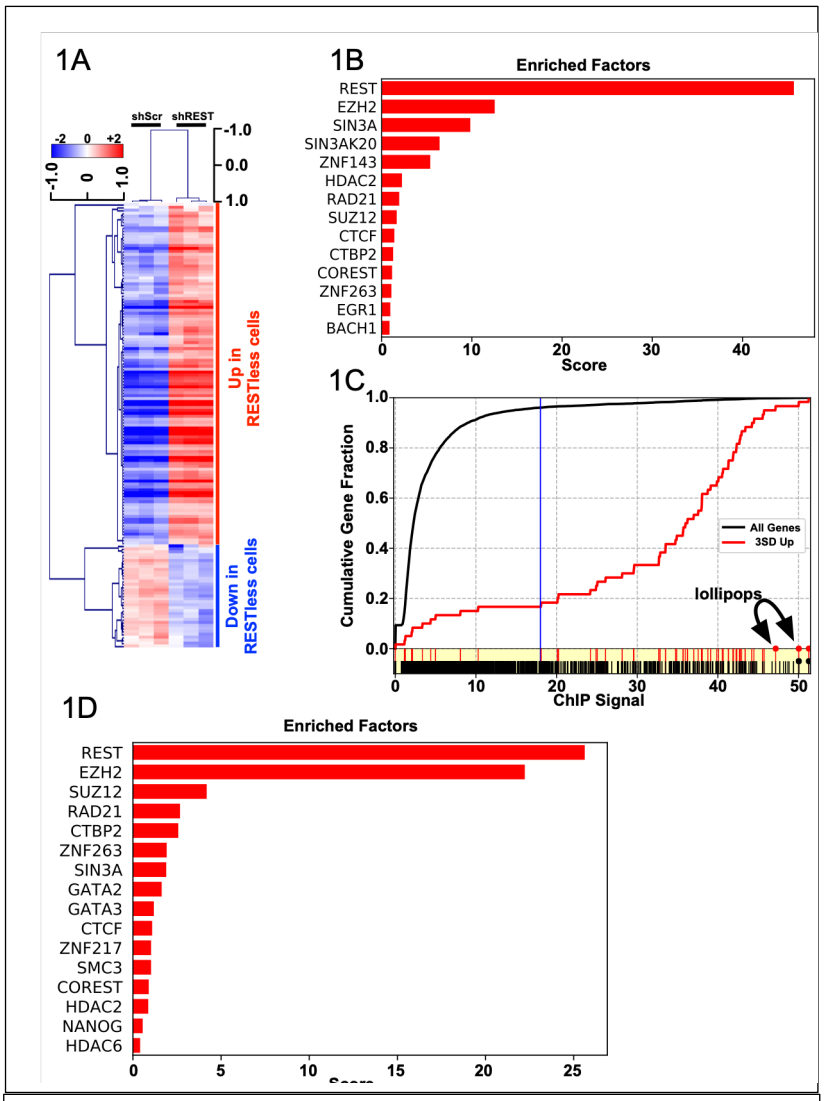

Figure 1. Analysis of MCF7 cells expressing shRNA against REST. **A**) Heatmap showing 119 up-regulated and 35 down-regulated genes by more than 3 standard deviations from the mean fold change at FDR< 0.05. **B**) MAGICACT summary output for up-regulated genes. **C**) Cumulative Distributions Functions of REST ChIP signals extracted from the MAGIC matrix for up-regulated genes (red) or all genes in the analysis (black). Black ticks along the bottom of the graph mark all genes in the analysis. Red ticks mark all up-regulated genes. Lollipops representing the 5%ile of highest ChIP signals in the up-regulated genes list (red) and the same number of highest ChIP signals in the whole analysis (black) are labelled as 'lollipops'. **D**) MAGICACT summary file using genes up-regulated 1.2x at FDR<0.05 in cells expressing shRNA against REST.

the population of REST ChIP values (black) and those associated with genes induced more than 3 standard deviations of the mean change (red).
Differentially expressed genes are often defined by a fold change and an associated False Discovery Rate. We asked whether defining up-regulated genes as those induced at least 1.2 fold with FDR<0.25 and raw p value <0.05 would alter the outcome of MAGICACT. Figure 1D shows that using this relaxed criterion for upregulation resulted in an expanded input list of 158 genes (supplemental table 8), but REST was still being called as the most significant Factor along with its cohort of cofactors.

For comparison, we used the 3SD gene list in oPossum. Although oPossum did predict REST as the most likely driver (z score = 83) (supplemental table 2), oPossum predicted 45 other transcription factors at z>1.65. Most of the factors have no known physical or functional association with REST. REST is a unique TF in that it has an extended binding motif of 23bp, which likely aided its identification by oPossum. Regardless, oPossum was not able to call any REST cofactors as drivers of the gene list.

Next we tested MAGICACT on The Cancer Genome Atlas (TCGA) Nature 2012(10) provisional breast cancer dataset. We extracted 233 Luminal-A and 79 Basal-like tumors. Luminal-A tumors are a subset of breast cancers defined by their robust expression of estrogen receptor alpha (ERα) and associated pioneer factors. They also express the progesterone and her2 receptors. Basal-like tumors are estrogen receptor negative and also lack progesterone and her2 receptors. This is a complex dataset with over 300 samples and heterogeneous tissue. We took all genes up-regulated more than 3SD from the mean fold change in Luminal-A tumors relative to Basal-like (203 genes) as the gene list for input into MAGICACT and 17814 expressed genes (supplemental table 10). MAGICACT called ERα (ESR1) as the highest scoring factor ($p=2.9x10^{-14}$, FDR = $5.0x10^{-13}$, Score = 8.8). Importantly, MAGICACT also called ERα pioneer factors as significant Factors at FDR<0.05 (FOXA1, FOXA2, GATA2 and GATA3)(11), known cofactors EZH2(12) and P300(13) (Fig.2 and supplemental table 3) and ZNF217(14). The same gene list yielded 46 TFs with oPossum at z <1.65 with ERα ranked at 41 (z=2.633). FOXA2 (z=23.2) and FOXA1 (z=20.1) were the top hits (supplemental table 4). Thus, whereas MAGICACT predicted ERα as a principal factor and identified its cohort

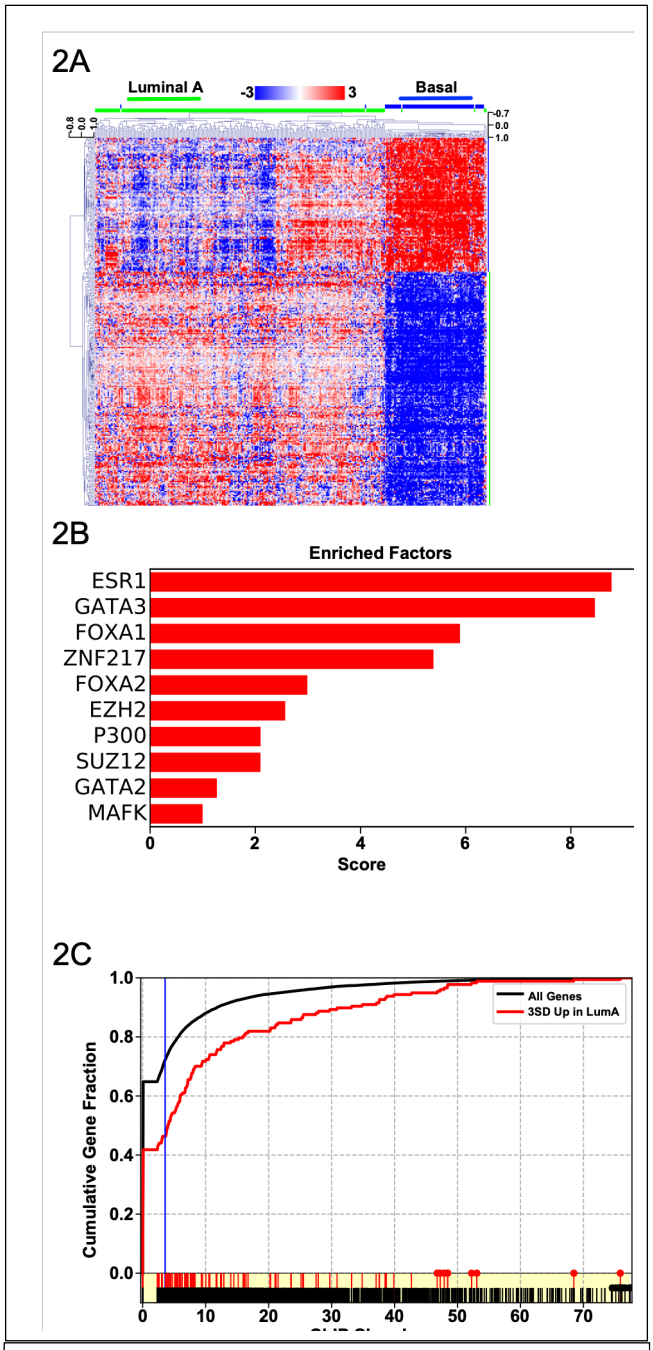

Figure 2. Analysis of Luminal-A and Basal like tumors in TCGA. **A**) Heatmap showing 203 up-regulated and 143 down-regulated genes by more than 3 standard deviations from the mean fold change at FDR < 0.05 across 233 Luminal-A and 79 Basal-like tumors **B**) MAGICACT summary output for up-regulated genes. **C**) Cumulative Distributions Functions of REST ChIP signals extracted from the MAGIC matrix for up-regulated genes (red) and 17814 background genes in the analysis (black).

of pioneer factors and cofactors in ER positive breast cancer, oPossum was unable to clearly underscore the estrogen receptor biology in this test.

MAGICACT was developed using ChIPseq data derived from immortalized or transformed human cancer cell lines and the above two examples utilize either an immortalized cell line or cancer tissue. To test whether MAGICACT could be used on data derived from non-cancerous, highly heterogeneous, rodent samples, we turned to non-malignant mouse brain tissue. CTCF is a TF that organizes long distance interactions in the genome. Sams et al demonstrated that elimination of CTCF in excitatory neurons (while sparing its expression in other neuron subtypes, astocytes and glia) results in defects in learning, memory and neuronal plasticity(15). RNA from whole hippocampus (i.e. all cell types in the formation) from WT and CTCF knock-out mice (n=3 per condition) was subjected to RNAseq and the transcriptome data was archived as GSE84175. We generated two gene lists from this dataset: genes up or down-regulated upon CTCF loss at least 1.2x with FDR<0.25 and raw p<0.05 (696 and 432 genes respectively – supplemental table 11) (Fig. 3A). Figure 3B and supplemental table 5 shows that the 3 highest scoring Factors targeting genes that are down-regulated in CTCF-KO hippocampii are CTCFL, ZNF143 and CTCF itself. The same list in oPOSSUM yielded ZNF354C as the top hit with CTCF ranked number 36 (z=3.8) (supplemental table 6). Interestingly, for the list of up-regulated genes, MAGICACT yielded MAFF and MAFK as top hits (Fig. 3C) which are associated with Locus Control Regions that require CTCF function(16).

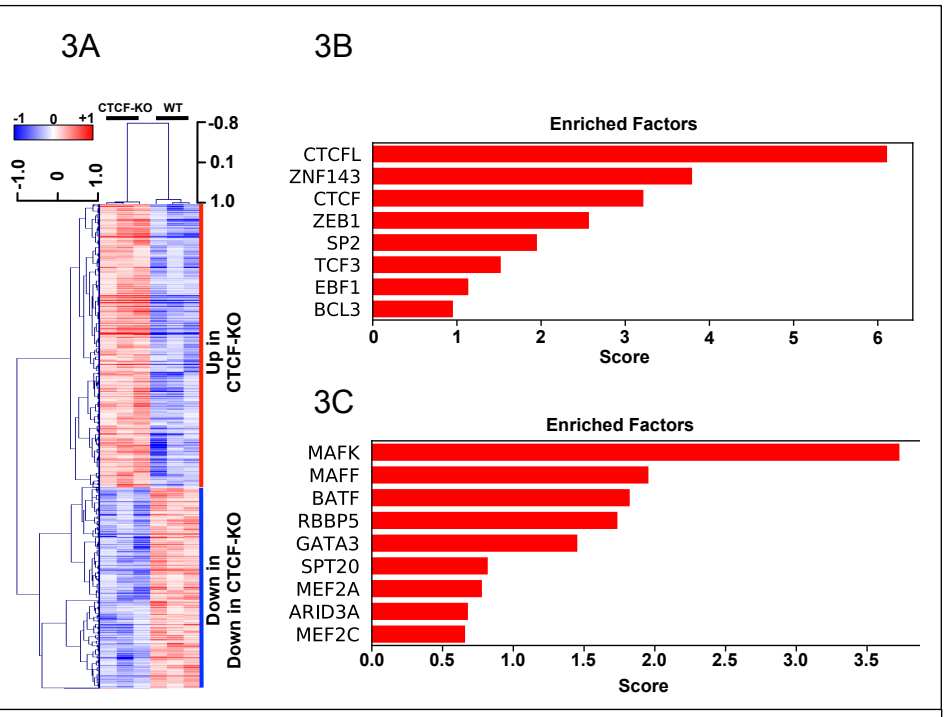

Figure 3. Analysis of WT and CTCF Knock-out hippocampal tissue (GSE84175) . **A**) Heatmap showing 696 up-regulated and 432 down-regulated genes changes at least 1.2x with FDR<0.25 and raw p<0.05 **B**) MAGICACT summary output for up-regulated genes. **C**) MAGICACT summary output for down-regulated genes.

In summary, MAGICACT was able to predict which TFs and cofactors drive gene changes in three different types of transcriptomic analysis: a cell line with or without REST, a tumor cohort with or without ERα and brain samples with or without CTCF.

Discussion.

We introduce MAGICACT as a standalone application that predicts transcription factors and cofactors that drive gene expression changes in transcriptomic experiments. In comparison to a commonly used resource, oPossum, MAGICACT predicted factors that were directly relevant to the biology behind the experiment. MAGICACT predicted both the transcription factors whose binding are enriched in the input list as well as cofactors that are consistent with the known biology of the cognate transcription factor.

MAGICACT was developed in response to an unmet need that would allow allocation of transcription factors and cofactors to lists of genes obtained through whole transcriptome experiments. A widely used approach - exemplified by oPossum - rely on searching for short transcription factor binding motifs in promoter proximal sequences. Using hypergeometric tests provides a likelihood of enrichment for the factor. This approach is

limited due to the small size of binding sites for most factors; a random 6bp sequence will be found approximately 750,000 times in the human genome. This approach also cannot predict cofactors as cofactors do not bind DNA directly.

An alternative approach relies on comparing a query list of genes with gene lists attributed to factors based on an arbitrary minimal signal in ChIPseq experiments. This approach does not take into account the fact that many factors have highly skewed binding profiles such that many genes will show some, low level of binding/chip signal. For example, extracting all REST ChIPseq tracks from wgEncodeRegTfbsClusteredV2.bed from the UCSC gemome browser (at http://hgdownload.soe.ucsc.edu/goldenPath/hg19/encodeDCC/wgEncodeRegTfbsClustered/) showed 27,386 binding sites across the human genome with 17,971 genes showing a detectable signal within 5Kb of the promoter (data not shown). This thwarts unbiased attempts to define genes as 'targets' or 'non-targets'. MAGICACT overcomes these shortfalls by utilizing complete ChIPseq profiles without parsing Factors into 'bound' and 'not bound' classes.

Testing MAGICACT on cells transfected with REST shRNA highlights the ability of the algorithm to not only identify transcription factors but also their associated cofactors. MAGICACT called REST as the highest scoring factor but also called SIN3A, HDAC2, CoREST and CtBP2, which are well-characterized REST cofactors(17-20). EZH2 and SUZ12 were also called as factors that preferentially bind up-regulated genes. Though there are reports of REST interacting with Polycomb(21), it is likely that Polycomb targets many of the same genes as REST but does so independently(22). In either case, using a well defined clonal cell line, MAGICACT calls Factors that are either cofactors for REST or co-regulate the same genes. Using gene lists derived from over 300 breast tumor transcriptomes, MAGICACT also correctly called ERα as the principal driver of gene induction in ER positive (Luminal-A) versus negative (Basal-like) tumors. ERα binds chromatin with pioneer factors and MAGICACT sucessfully called GATA2, GATA3, FOXA1 and FOXA2(11). P300 and EZH2 are ERα cofactors and were called by MAGICACT(12,13).

In comparison, oPossum was able to identify REST as the principal factor in the first analysis by using Fisher tests of the input list against lists of genes defined as transcription factor targets by proximity to sequence motifs. As expected, oPossum was unable to call any cofactors due to an absence of binding motifs for this class of nuclear protein. With the breast cancer samples, oPossum ranked ERα 41 out of 116 candidate factors. Thus, using oPossum, a naïve researcher studying transcriptomes from Luminal-A and Basal-like tumors would have had difficulty identifying estrogen receptor as a principal factor in the biology of this disease. However, the ERα pioneer factors FOXA1 and FOXA2 were ranked #1 and 2 by oPossum, which would hint at a role for steroid receptors in the disease. Nevertheless, MAGICACT highlighted estrogen receptor biology as a primary driver of transcriptional differences between the 2 groups of tumors.

ENCODE hosts ChipSeq tracks derived from transformed or immortal cell lines and human embryonic stem cells. The MAGIC Matrix incorporates the highest ChIP signal observed across all cell lines within a gene domain for a given factor. This approach was chosen (rather than taking the mean signal for example) because MAGICACT is a discovery platform and taking the maximum value is an attempt to maximize the chance that that particular site may be bound in other systems. A particular concern was that generating a matrix from homogenous cell lines may preclude MAGICACT's use in analysis of transcriptome data derived from in vivo samples that are not immortalized and likely heterogeneous. However, testing of MAGICACT on GSE84175 demonstrates that RNAseq data derived from control and experimental whole brain extract is handled appropriately. In GSE84175, CTCF was deleted in a subset of post-mitotic cells (eliminated from excitatory neurons but retained in all other neuronal subtypes) (15,23). RNA was extracted from whole hippocampus, which included glia and astrocytes. MAGICACT correctly called CTCF and CTCFL as the factors most likely to drive gene changes in the experiment. Interestingly, ZNF143 was also in the top 3 hits and EBF1 was also called. This is in keeping with the recent findings of Mourad and Cuvier(23) showing that CTCF and ZNF143

coordinate chromatin border domain formation. Also, Zhang et al demonstrated the role of EBF1 in controlling higher-order chromatin structure(24). Thus, MAGICACT was able to call CTCF/L as well as a known associated factor when provided transcriptome data from a highly heterogeneous tissue where only a small subset of cells were altered in the experiment.

Another popular resource for Factor prediction is Enrichr (25) that accepts input lists and predicts factors based on Chea and ENCODE as well as literature mining. Enricher works by taking the top 2000 ChIPed genes for each Factor in each cell line in ENCODE as the 'target list' for that Factor. Fisher tests are then performed to assess association between the query list and the 2000 targets. Enrichr was accurately able to predict REST, ESR1 and CTCF in the 3 examples used in this manuscript (data not shown). The noticable difference between MAGICACT and Enrichr was in the identification of cofactors. When genes which are upregulated in RESTless MCF7 cells were used as input, Enrichr correctly predicted REST as the driver. However, it was unable to call SIN3A or CoREST using ENCODE and ChEA Consensus from ChIP-X whereas MAGICACT was able to do so. Similarly, when genes present in Luminal-A tumors was used as input, Enrichr predicted ESR1 as a driver but, unlike MAGICACT, failed to call the GATA pioneer factors. Thus MAGICACT was able to more clearly define the biology behind the experiment than Enrichr. It is possible that 2 issues confound a more robust prediction from Enrichr. Firstly, the user cannot enter a backgound gene list and so Fisher tests will inflate significance of association of unrelated factors. Secondly, the arbitrary cutoff of 2000 genes prevents incorporation of weaker Factor targets that may still be biologically important.

A clear limitation of the current input matrix is that only factors with peaks within 5Kb of the gene body were assigned to a gene: it is clear that a more complete method will take into account distal enhancers that have major roles in global gene regulation. However, despite the absence of distal enhancer information in the MAGIC Matrix, the above three examples show the utility of MAGICACT and MAGIC Matrix using only gene-proximal factor binding. Another limitation is the reliance on a matrix comprised of 161 ENCODE ChIPseq tracks despite there being over 1000 proteins in the human genome that reside in the nucleus. However, the algorithm can accept user defined ChIP seq data and so can evolve as more tracks become available. Users can add their own ChIPseq data to the matrix. To do so, the user would find all ChIPseq peaks within 5Kb of all gene bodies in the MAGIC Matrix and pick the highest to assign to that gene. The sum of all peaks so defined would be normalized to sum to 100,000 and added to the MAGIC matrix as a column. Alternatively, the column can be used as the input matrix by itself. The single column matrix would be placed in the MAGICACT application directory and named MAGIC_Matrix.txt; MAGICACT will use the new matrix in place of the original.

In summary, MAGICACT is an application for identifying factors that preferentially bind lists of genes with a substantially reduced false positive rate compared to currently available methods.

Figure Legends.

Figure 1. Analysis of MCF7 cells expressing shRNA against REST. A) Heatmap showing 119 up-regulated and 35 down-regulated genes by more than 3 standard deviations from the mean fold change at FDR < 0.05. B) MAGICACT summary output for up-regulated genes. C) Cumulative Distributions Functions of REST ChIP signals extracted from the MAGIC matrix for up-regulated genes (red) or all genes in the analysis (black). Black ticks along the bottom of the graph mark all genes in the analysis. Red ticks mark all up-regulated genes. Lollipops representing the 95%ile of highest ChIP signals in the up-regulated genes list (red) and the same number of highest ChIP signals in the whole analysis (black) are labelled as 'lollipops'. D) MAGICACT summary file using genes up-regulated 1.2x at FDR<0.05 in cells expressing shRNA against REST.

Figure 2. Analysis of Luminal-A and Basal like tumors in TCGA. A) Heatmap showing 203 up-regulated and 143 down-regulated genes by more than 3 standard deviations from the mean fold change at FDR

< 0.05 across 233 Luminal-A and 79 Basal-like tumors B) MAGICACT summary output for up-regulated genes. C) Cumulative Distributions Functions of REST ChIP signals extracted from the MAGIC matrix for up-regulated genes (red) and 17814 background genes in the analysis (black).

Figure 3. Analysis of WT and CTCF Knock-out hippocampal tissue (GSE84175) . A) Heatmap showing 696 up-regulated and 432 down-regulated genes changes at least 1.2x with FDR<0.25 and raw p<0.05 B) MAGICACT summary output for up-regulated genes. C) MAGICACT summary output for down-regulated genes.

Conflicts of Interest.
The author has no conflicts of interest to declare.

Acknowledgements.
I would like to thank Ray Dingledine, Emory University for critical insights into the statistics behind MAGICACT and generation of MAGICACT_Windows. I thank John Svaren and Steve Goldstein (University of Wisconsin at Madison) for helpful edits of the manuscript. I am grateful to Caroline Alexander, University of Wisconsin-Madison for aiding, encouraging and cojoling the submission of this manuscript. I would like to thank Queen because 'It's a Kind of Magic'.

Bibliography.

1. Cancer Genome Atlas Research, N., Weinstein, J.N., Collisson, E.A., Mills, G.B., Shaw, K.R., Ozenberger, B.A., Ellrott, K., Shmulevich, I., Sander, C. and Stuart, J.M. (2013) The Cancer Genome Atlas Pan-Cancer analysis project. *Nat Genet*, **45**, 1113-1120.
2. Biggin, M.D. (2011) Animal transcription networks as highly connected, quantitative continua. *Dev Cell*, **21**, 611-626.
3. Kwon, A.T., Arenillas, D.J., Worsley Hunt, R. and Wasserman, W.W. (2012) oPOSSUM-3: advanced analysis of regulatory motif over-representation across genes or ChIP-Seq datasets. *G3 (Bethesda)*, **2**, 987-1002.
4. Meyer, K., Albaugh, B., Schoenike, B. and Roopra, A. (2015) Type 1 Insulin-Like Growth Factor Receptor/Insulin Receptor Substrate 1 Signaling Confers Pathogenic Activity on Breast Tumor Cells Lacking REST. *Mol Cell Biol*, **35**, 2991-3004.
5. Bolstad, B.M., Irizarry, R.A., Astrand, M. and Speed, T.P. (2003) A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics*, **19**, 185-193.
6. Nielsen, T.O., Parker, J.S., Leung, S., Voduc, D., Ebbert, M., Vickery, T., Davies, S.R., Snider, J., Stijleman, I.J., Reed, J. *et al.* (2010) A comparison of PAM50 intrinsic subtyping with immunohistochemistry and clinical prognostic factors in tamoxifen-treated estrogen receptor-positive breast cancer. *Clin Cancer Res*, **16**, 5222-5232.
7. Subramanian, A., Tamayo, P., Mootha, V.K., Mukherjee, S., Ebert, B.L., Gillette, M.A., Paulovich, A., Pomeroy, S.L., Golub, T.R., Lander, E.S. *et al.* (2005) Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci U S A*, **102**, 15545-15550.
8. Gunsalus, K.T., Wagoner, M.P., Meyer, K., Potter, W.B., Schoenike, B., Kim, S., Alexander, C.M., Friedl, A. and Roopra, A. (2012) Induction of the RNA regulator LIN28A is required for the growth and pathogenesis of RESTless breast tumors. *Cancer Res*, **72**, 3207-3216.
9. Wagoner, M.P., Gunsalus, K.T., Schoenike, B., Richardson, A.L., Friedl, A. and Roopra, A. (2010) The transcription factor REST is lost in aggressive breast cancer. *PLoS Genet*, **6**, e1000979.
10. The Cancer Genome Atlas, N., Koboldt, D.C., Fulton, R.S., McLellan, M.D., Schmidt, H., Kalicki-Veizer, J., McMichael, J.F., Fulton, L.L., Dooling, D.J., Ding, L. *et al.* (2012) Comprehensive molecular portraits of human breast tumours. *Nature*, **490**, 61.
11. Zaret, K.S. and Carroll, J.S. (2011) Pioneer transcription factors: establishing competence for gene expression. *Genes Dev*, **25**, 2227-2241.
12. Hwang, C., Giri, V.N., Wilkinson, J.C., Wright, C.W., Wilkinson, A.S., Cooney, K.A. and Duckett, C.S. (2008) EZH2 regulates the transcription of estrogen-responsive genes through association with REA, an estrogen receptor corepressor. *Breast Cancer Res Treat*, **107**, 235-242.
13. Shang, Y., Hu, X., DiRenzo, J., Lazar, M.A. and Brown, M. (2000) Cofactor dynamics and sufficiency in estrogen receptor-regulated transcription. *Cell*, **103**, 843-852.
14. Frietze, S., O'Geen, H., Littlepage, L.E., Simion, C., Sweeney, C.A., Farnham, P.J. and Krig, S.R. (2014) Global analysis of ZNF217 chromatin occupancy in the breast cancer cell genome reveals an association with ERalpha. *BMC Genomics*, **15**, 520.

15. Sams, D.S., Nardone, S., Getselter, D., Raz, D., Tal, M., Rayi, P.R., Kaphzan, H., Hakim, O. and Elliott, E. (2016) Neuronal CTCF Is Necessary for Basal and Experience-Dependent Gene Regulation, Memory Formation, and Genomic Structure of BDNF and Arc. *Cell Rep*, **17**, 2418-2430.
16. Sawado, T., Igarashi, K. and Groudine, M. (2001) Activation of beta-major globin gene transcription is associated with recruitment of NF-E2 to the beta-globin LCR and gene promoter. *Proc Natl Acad Sci U S A*, **98**, 10226-10231.
17. Andres, M.E., Burger, C., Peral-Rubio, M.J., Battaglioli, E., Anderson, M.E., Grimes, J., Dallman, J., Ballas, N. and Mandel, G. (1999) CoREST: a functional corepressor required for regulation of neural-specific gene expression. *Proc Natl Acad Sci U S A*, **96**, 9873-9878.
18. Garriga-Canut, M., Schoenike, B., Qazi, R., Bergendahl, K., Daley, T.J., Pfender, R.M., Morrison, J.F., Ockuly, J., Stafstrom, C., Sutula, T. *et al.* (2006) 2-Deoxy-D-glucose reduces epilepsy progression by NRSF-CtBP-dependent metabolic regulation of chromatin structure. *Nat Neurosci*, **9**, 1382-1387.
19. Huang, Y., Myers, S.J. and Dingledine, R. (1999) Transcriptional repression by REST: recruitment of Sin3A and histone deacetylase to neuronal genes. *Nat Neurosci*, **2**, 867-872.
20. Roopra, A., Sharling, L., Wood, I.C., Briggs, T., Bachfischer, U., Paquette, A.J. and Buckley, N.J. (2000) Transcriptional repression by neuron-restrictive silencer factor is mediated via the Sin3-histone deacetylase complex. *Mol Cell Biol*, **20**, 2147-2157.
21. Dietrich, N., Lerdrup, M., Landt, E., Agrawal-Singh, S., Bak, M., Tommerup, N., Rappsilber, J., Sodersten, E. and Hansen, K. (2012) REST-mediated recruitment of polycomb repressor complexes in mammalian cells. *PLoS Genet*, **8**, e1002494.
22. McGann, J.C., Oyer, J.A., Garg, S., Yao, H., Liu, J., Feng, X., Liao, L., Yates, J.R., 3rd and Mandel, G. (2014) Polycomb- and REST-associated histone deacetylases are independent pathways toward a mature neuronal phenotype. *Elife*, **3**, e04235.
23. Mourad, R. and Cuvier, O. (2016) Computational Identification of Genomic Features That Influence 3D Chromatin Domain Formation. *PLoS Comput Biol*, **12**, e1004908.
24. Zhang, X., Branciamore, S., Gogoshin, G., Rodin, A.S. and Riggs, A.D. (2017) Analysis of high-resolution 3D intrachromosomal interactions aided by Bayesian network modeling. *Proc Natl Acad Sci U S A*, **114**, E10359-E10368.
25. Chen, E.Y., Tan, C.M., Kou, Y., Duan, Q., Wang, Z., Meirelles, G.V., Clark, N.R. and Ma'ayan, A. (2013) Enrichr: interactive and collaborative HTML5 gene list enrichment analysis tool. *BMC Bioinformatics*, **14**, 128.