ELMER v.2: An R/Bioconductor package to reconstruct gene regulatory networks from DNA methylation and transcriptome profiles

Tiago C Silva[1,2] Simon G Coetzee[2] Lijing Yao[3] Nicole Gull[2] Dennis J Hazelett[2] Houtan Noushmehr[5] De-Chen Lin[4,*] Benjamin P Berman[2,4,*]

1 Department of Genetics, Ribeirão Preto Medical School, University of São Paulo, Ribeirão Preto, Brazil,
2 Center for Bioinformatics and Functional Genomics, Cedars-Sinai Medical Center, Los Angeles, CA, USA
3 Roche Sequencing Solutions, Belmont, CA, USA
4 Department of Medicine, Cedars-Sinai Medical Center, Los Angeles, California, USA and
5 Department of Neurosurgery, Henry Ford Hospital, Detroit, MI, USA

*** dchlin11@gmail.com or Benjamin.Berman@csmc.edu**

Abstract

Motivation: DNA methylation can be used to identify functional changes at transcriptional enhancers and other cis-regulatory modules (CRMs) in tumors and other primary disease tissues. Our R/Bioconductor package *ELMER* (Enhancer Linking by Methylation/Expression Relationships) provides a systematic approach that reconstructs gene regulatory networks (GRNs) by combining methylation and gene expression data derived from the same set of samples.

Results: We present new *ELMER* version that provides a new *Supervised* analysis mode, which uses pre-defined sample groupings and can identify additional Master Regulators, such as *KLF5* in basal-like breast cancer.

Availability: *ELMER* (v2.0) is available as an R/Bioconductor package at `http://bioconductor.org/packages/ELMER/` with auxiliary data at `http://bioconductor.org/packages/ELMER.data/`.

Contact: dchlin11@gmail.com or Benjamin.Berman@csmc.edu

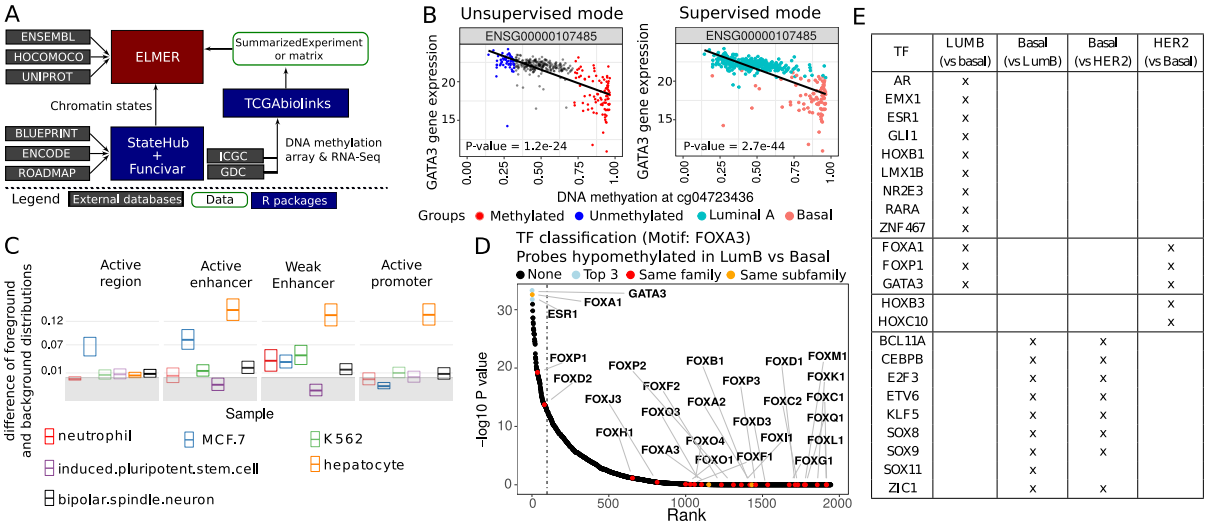
Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Motivated by our identification of transcription factor binding sites (TFBSs), enhancers, and other cis-regulatory modules (CRMs) from primary tumor DNA methylation data (Berman et al., 2012), we developed the R/Bioconductor *ELMER* (Enhancer Linking by Methylation/Expression Relationships) package, which infers regulatory element landscapes and transcription factor networks from cancer methylomes (Yao et al., 2015). Here, we have implemented a number of important new features and changes summarized in Supplementary Table S1. Amongst the most important are: (i) use of the gold-standard MultiAssayExperiment (MAE) class to encapsulate all experimental data types, (ii) integration with our TCGABiolinks (Colaprico et al., 2015) package to import data and define cohorts from Genomic Data Commons (Grossman et al., 2016) (iii) integration with our TCGABiolinksGUI (Silva et al., 2017) to allow ELMER runs to be performed in a web-based environment, (iv) all output integrated into a single interactive HTML file with all tables and figures, (v) annotation of cell-type specific chromatin context for all non-coding elements identified, and (vi) a new *Supervised* mode where changes are identified between two user-defined groups. Importantly, this final feature allows application of ELMER to arbitrary disease datasets, rather than the tumor/normal comparisons supported by ELMER version 1. Here, we describe these new ELMER v. 2.0 features, and show how they can be used to reveal new biological insights.

2 System and methods

In this section, we focus on key aspects that have been added or improved. For a full description of methods and code architecture, see the Supplemental Methods and Note.



TF	LUMB (vs basal)	Basal (vs LumB)	Basal (vs HER2)	HER2 (vs Basal)
AR	x			
EMX1	x			
ESR1	x			
GLI1	x			
HOXB1	x			
LMX1B	x			
NR2E3	x			
RARA	x			
ZNF467	x			
FOXA1	x			x
FOXP1	x			x
GATA3	x			x
HOXB3				x
HOXC10				x
BCL11A		x	x	
CEBPB		x	x	
E2F3		x	x	
ETV6		x	x	
KLF5		x	x	
SOX8		x	x	
SOX9		x	x	
SOX11		x		
ZIC1		x	x	

Figure 1. **A** ELMER interactions. **B** Groups U and M definition in *(un)supervised* mode. **C** BRCA supervised analysis: Candidate MRs. **D** Chromatin state enrichment analysis: enrichment for enhancer active region, weak enhancer and active promoter region for MCF-7 cell. **E** TFs ranking for the *FOXA3* motif. The top 3 most anti-correlated TFs (*FOXA1*, *GATA3* and *ESR1*), and all TF classified by TFClass database in the same (sub)family are highlighted.

Supervised vs Unsupervised mode

ELMER is designed to identify differences between two sets of samples within a given dataset. In original version (Yao et al., 2015), the first step - identification of differentially methylated CpG probes (DMCs) - was hard-coded to identify DMCs between non-cancer vs. cancer samples, and the subsequent step was *unsupervised*, identifying changes within any subset of tumors. In ELMER 2.0, we generalize these strategies so that they are applicable to any paired dataset, including disease vs. healthy tissue for any disease type.

In the *Unsupervised* mode, as in ELMER 1.0, for each comparison samples from two most extreme quintiles based on DNA methylation levels were used, generating both the M (methylated) and the U (unmethylated) groups. In the new *Supervised* mode, the U and M groups are defined strictly by sample group labels, and all samples in each group are used. The *Supervised* mode can greatly increase statistical power (Fig. 1B).

Characterization of chromatin state context of enriched probes

To provide possible functional interpretation of the regulatory elements identified by *ELMER*, we perform chromatin state enrichment analysis of identified probes, using `statehub.org` data (Coetzee et al., 2018). Enrichment of chromatin states is calculated against a background model that uses the distal probe set from which the candidate probes are drawn from (Fig. 1C).

Motif enrichment analysis

HOCOMOCO v11 (Kulakovskiy et al., 2016) TF binding models were used as input for HOMER (Heinz et al., 2010) to find motif occurrences in a $\pm 250bp$ region around each probe from EPIC and HM450 arrays. A motif enrichment analysis using Fisher's exact test and Benjamini-Hochberg multiple hypothesis testing correction is performed.

Identification of master regulator TFs

For each enriched motif, *ELMER* takes the mean DNA methylation of all distal probes (in significant probe-gene pairs) that contain that motif occurrence (within a $\pm 250bp$ region) and compares this mean DNA methylation to the expression of each gene annotated as a human TF.

3 Use Case

We performed *ELMER* analysis on TCGA BRCA (Breast Invasive Carcinoma) data retrieved directly from the GDC server. Use Case 1 compares tumor samples to normal samples using the *Unsupervised* analysis mode, while Use Case 2 compares known BRCA molecular subtypes (Her2, Luminal A, Luminal B and Basal-like) using the *Supervised* mode. The HTML output report containing the code and results of both analysis are available as supplemental material and at `http://bit.ly/ELMER_reports`.

Use Case 1: Breast Invasive Carcinoma (Unsupervised mode)

For the most highly enriched motif *FOXA3*, the master regulator analysis identified *FOXA1* as the top candidate among all TFs in the human genome (Fig. 1D), with the collaborating factors *GATA3* and *ESR1* being discovered as the next best candidates (Supplementary Fig. 14).

Use Case 2: BRCA molecular subtypes analysis (Supervised mode)

Fig. 1E summarizes the candidate master regulator TFs (MRs) identified. For the Basal-like group, regulatory TF candidate identified are the *SOX11* and *KLF5*. Importantly, their functional relevance in basal-like BRCA was recently described (Ben-Porath et al., 2008; Shepherd et al., 2016).

4 Conclusion

ELMER v2.0 added the *Supervised* analysis mode, which can be used to identify novel cancer subtype-specific Master Regulators. For our basal-like breast cancer analysis, known and novel MR candidates were identified including *SOX11* and *KLF5*.

Funding

The project was funded by the Cedars-Sinai's Samuel Oschin Comprehensive Cancer Institute, by FAPESP (2016/01389-7 to T.C.S. & H.N. and 2015/07925-5 to H.N.), by NCI ITCR program (1U01CA184826 to B.P.B., D.J.H & S.G.C), and by GDAN (1U24CA210969 to B.P.B & T.C.S, R01CA190182 to D.J.H).

Competing interests

No competing interests were disclosed

References

Ben-Porath, I. et al. (2008). An embryonic stem cell–like gene expression signature in poorly differentiated aggressive human tumors. *Nature genetics*, 40(5):499–507.

Berman, B. P. et al. (2012). Regions of focal dna hypermethylation and long-range hypomethylation in colorectal cancer coincide with nuclear lamina-associated domains. *Nature genetics*, 44(1):40–46.

Coetzee, S. et al. (2018). Statehub-statepaintr: rapid and reproducible chromatin state evaluation for custom genome annotation. *F1000Research*, 7(214).

Colaprico, A., Silva, T. C., et al. (2015). Tcgabiolinks: an r/bioconductor package for integrative analysis of tcga data. *Nucleic acids research*, page gkv1507.

Grossman, R. L. et al. (2016). Toward a shared vision for cancer genomic data. *New England Journal of Medicine*, 375(12):1109–1112.

Heinz, S. et al. (2010). Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and b cell identities. *Molecular cell*, 38(4):576–589.

Kulakovskiy, I. V. et al. (2016). Hocomoco: expansion and enhancement of the collection of transcription factor binding sites models. *Nucleic acids research*, 44(D1):D116–D125.

Shepherd, J. H. et al. (2016). The sox11 transcription factor is a critical regulator of basal-like breast cancer growth, invasion, and basal-like gene expression. *Oncotarget*.

Silva, T. C. et al. (2017). Tcgabiolinksgui: A graphical user interface to analyze gdc cancer molecular and clinical data. *bioRxiv*.

Yao, L. et al. (2015). Inferring regulatory element landscapes and transcription factor networks from cancer methylomes. *Genome biology*, 16(1):105.

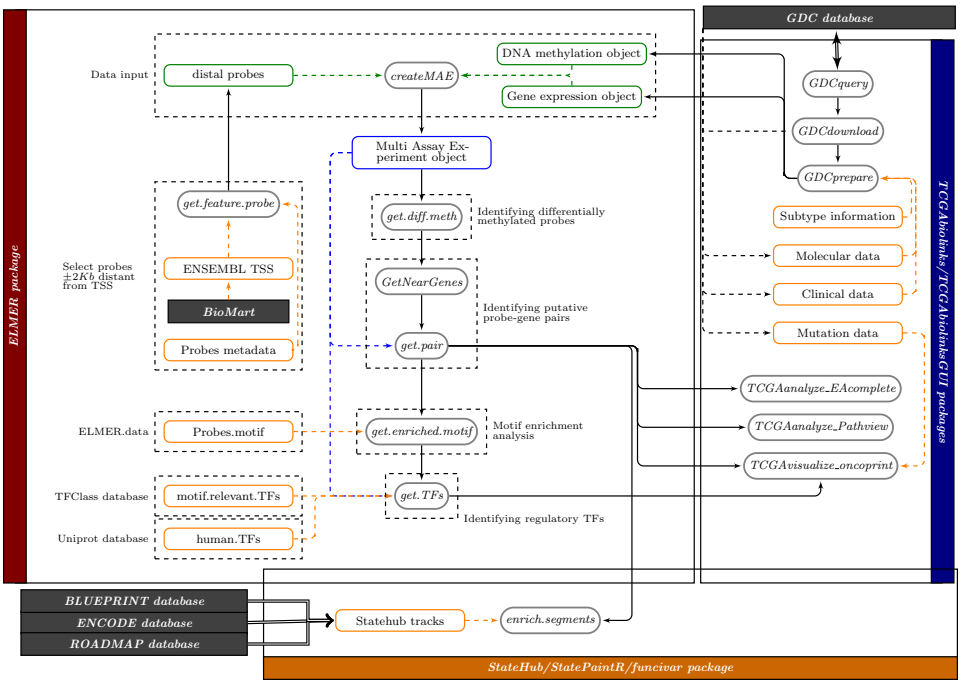
Supplementary Methods and Notes for *ELMER* v.2: An R/Bioconductor package to reconstruct gene regulatory networks from DNA methylation and transcriptome profiles

Introduction

In addition to the details below, a complete HTML output report for the two runs described in the *Use case* Section is available at `http://bit.ly/ELMER_reports`. This document contains all source code, parameters used, Methods descriptions, output tables, and output plots.

ELMER workflow

The complete ELMER workflow is shown in Supplementary Fig. 1.



Supplementary Fig 1. ELMER workflow: ELMER receives as input a DNA methylation array object and a gene expression object (matrices or SummarizedExperiment objects) and a Genomic Ranges (GRanges) object with distal probes to be used as filter which can be retrieved using the *get.feature.probe* function. The function *createMAE* will create a Multi Assay Experiment object keeping only samples that have both DNA methylation and gene expression data. Genes will be mapped to genomic position and annotated using ENSEMBL database (Aken et al., 2016), while for probes it will add annotation from Zhou et al. (http://zwdzwd.github.io/InfiniumAnnotation) . This MAE object will be used as input to the next analysis functions. First, it identifies differentially methylated probes followed by the identification of their nearest genes (10 upstream and 10 downstream) through the *get.diff.meth* and *GetNearGenes* functions respectively. For each probe, it will verify if any of the nearby genes were affected by its change in the DNA methylation level and a list of gene and probes pairs will be outputted from *get.pair* function. For the probes in those pairs, it will search for enriched regulatory Transcription Factors motifs with the *get.enriched.motif* function. Finally, the enriched motifs will be correlated with the level of the transcription factor through the *get.TFs* function. In the figure green Boxes represents user input data, blue boxes represent output object, orange boxes represent auxiliary pre-computed data and gray boxes are functions.

Main differences between ELMER old version (v.1) and the new version (v.2)

The main differences between ELMER old version (v.1) and the new version (v.2) is summarized in the Supplementary Table S1.

Table S1. Main differences between ELMER old version (v.1) and the new version (v.2)

Features	ELMER Version 1	ELMER Version 2
Primary data structure	mee object (custom data structure)	MAE object (Bioconductor data structure)
Auxiliary data	Manually created	Programmatically created
Number of human TFs	1,982	2,014 (UniProt database)
Number of TF motifs	91	771 (HOCOMOCO v11 database)
TF classification	78 families	82 families and 331 subfamilies (TFClass database, HOCOMOCO)
Analysis performed	Normal vs tumor samples	Group 1 vs group 2
Statistical grouping	Unsupervised only	Unsupervised or supervised using labeled groups
TCGA data source	The Cancer Genome Atlas (TCGA) (not available)	The NCI's Genomic Data Commons (GDC)
Genome of reference	GRCh37 (hg19)	GRCh37 (hg19)/GRCh38 (hg38)
DNA methylation platforms	HM450	EPIC and HM450
Graphical User Interface (GUI)	None	TCGAbiolinksGUI
Automatic report	None	HTML summarizing results
Annotations	None	StateHub

Organization of data as a *MultiAssayExperiment* object

To facilitate the analysis of experiments and studies with multiple samples the Bioconductor team created the *SummarizedExperiment* class (Huber et al., 2015), a data structure able to store data and metadata for a single experiment but not for data spanning several experiments for the same sample. To overcome this problem, recently, the MultiAssay SIG (Special Interest Group) created the MultiAssayExperiment class (Ramos et al., 2017) a data structure to manage and preprocess multiple assays for integrated genomic analysis. This data structure is now an input for all main functions of *ELMER* and can be generated by the *createMAE* function.

To perform *ELMER* analyses, we populate a *MultiAssayExperiment* with a DNA methylation matrix or *SummarizedExperiment* object from HM450K or EPIC platform; a gene expression matrix or SummarizedExperiment object for the same samples; a matrix mapping DNA methylation samples to gene expression samples; and a matrix with sample metadata (i.e. clinical data, molecular subtype, etc.). TCGA or other GDC data can be imported by TCGAbiolinks (Supplementary Figure 1), in which case the necessary data structures will be automatically created. Based on the genome of samples selected, metadata for the DNA methylation probes, such as genomic coordinates, will be added from (Zhou et al., 2016); and metadata for gene annotation is added from the ENSEMBL database (Yates et al., 2015) using biomaRt (Durinck et al., 2009). Use of these standardized import packages allows *ELMER 2.0* to take advantage of all current datasets. For instance, TCGABiolinks will soon be able to read from the International Cancer Genome Consortium (ICGC) repository, and similar importers can be written for other disease databanks.

If using non-TCGA data, the matrix with sample metadata should be provided with at least a column with a patient identifier and another one identifying its group which will be used for analysis, if samples in the methylation and expression matrices are not ordered and with same names, a matrix mapping for each patient identifier their DNA methylation samples and their gene expression samples

should be provided to the *createMAE* function.

Selecting distal probes

Probes from HumanMethylationEPIC (EPIC) array and Infinium HumanMethylation450 (HM450) array are removed from the analysis if they have either internal SNPs close to the $3'$ end of the probe; non-unique mapping to the bisulfite-converted genome; or off-target hybridization due to partial overlap with non-unique elements (Zhou et al., 2017).

Probe metadata information is available as *ELMER.data* package, populated from the source file at `http://zwdzwd.github.io/InfiniumAnnotation` (Zhou et al., 2017).

Also, for the analysis of distal elements, probes located in regions of $\pm 2kb$ around transcription start sites (TSSs) were removed.

Identification of differentially methylated CpGs (DMCs)

ELMER is designed to identify differences between two sets of samples within a given dataset. In Yao et al., the first step (identification of DMCs) was hard-coded to identify DMCs between non-cancer vs. cancer samples, and the subsequent step was *unsupervised*, identifying changes within any subset of tumors. In ELMER 2.0, we generalize these strategies so that they are applicable to any paired dataset, including disease vs. healthy tissue for any disease type, untreated vs. treated samples, etc. We now support two modes, with the *Unsupervised* based on the method of Yao et al.. Here, the user defines Group 1 and Group 2 samples, but an assumption is made that only a subset of samples differs between the two groups. By default, this subset includes the most extreme 20% of samples within the group. While the 20% is an input parameter that can be changed to any value, this was the setting used in Yao et al., which was able to identify molecular cancer subtypes such as ER-positive tumors within the complete set of TCGA breast cancer cases. The new mode is the *Supervised* mode, in which *all* available samples from each group are used. This mode should be used when pre-determined phenotypes or molecular subtypes are known in advance, such as the treated vs. untreated case. The advantage is that this greatly increases statistical power because of all samples from each group. This can be extremely important, given the large burden of multiple hypothesis testing involved in ELMER.

The first step is the identification of differentially methylated CpGs (DMCs). In the *Supervised* mode, we compare the DNA methylation level of each distal CpG for *all* samples in Group 1 compared to all samples Group 2, using an unpaired one-tailed t-test. In the *Unsupervised* mode, the samples of each group (Group 1 and Group 2) are ranked by their DNA methylation beta values for the given probe, and those samples in the lower quintile (20% samples with the lowest methylation levels) of each group are used to identify if the probe is hypomethylated in Group 1 compared to Group 2. The reverse applies for the identification of hypermethylated probes. It is important to highlight that in the *Unsupervised* mode, each probe selected may be based on a different subset the samples, and thus probe sets from multiple molecular subtypes may be represented. In the *supervised* mode, all tests are based on the same set of samples.

The 20% is a parameter to the *diff.meth* function called *minSubgroupFrac*. For the unsupervised analysis, this is set to 20% as in Yao et al. (2015), because we wanted to be able to detect a specific molecular subtype among samples; these subtypes often make up only a minority of samples, and 20% was chosen as a lower bound for the purposes of statistical power (high enough sample numbers to yield t-test p-values that could overcome multiple hypotheses corrections, yet low enough to be able to capture changes in individual molecular subtypes occurring in 20% or more of the cases.) This number can be set as an input to the *diff.meth* function and should be tuned based on sample sizes in individual studies. In the *Supervised* mode, where the comparison groups are implicit in the sample set and labeled, the *minSubgroupFrac* parameter is set to 100%. An example would be a cell culture experiment with 5 replicates of the untreated cell line, and another 5 replicates that include an experimental treatment.

To identify hypomethylated DMCs, a one-tailed t-test is used to rule out the null hypothesis: $\mu_{group1} \geq \mu_{group2}$, where μ_{group1} is the mean methylation within the lowest group 1 quintile (or another percentile as specified by the *minSubgroupFrac* parameter) and μ_{group2} is the mean within the lowest group 2 quintile. Raw p-values are adjusted for multiple hypothesis testing using the Benjamini-Hochberg method (Benjamini and Hochberg, 1995), and probes are selected when they had adjusted p-value less than 0.01 (which can be configured using the *pvalue* parameter). For additional stringency, probes are only selected if the methylation difference: $\Delta = \mu_{group1} - \mu_{group2}$ was greater than 0.3. The same method is used to identify hypermethylated DMCs, except we use the *upper* quintile, and the opposite tail in the t-test is chosen.

Identification of putative target gene(s)

For each differentially methylated distal probe (DMC), the closest 10 upstream genes and the closest 10 downstream genes are tested for inverse correlation between methylation of the probe and expression of the gene, which is the same basic strategy employed in ELMER version 1. However, we now import all gene annotations programmatically using the Biomart (Durinck et al., 2005, 2009) package. This allows easy extensibility to use any annotations desired (our default uses Ensembl annotations).

This step also differs between the *Supervised* and *Unsupervised* modes. In the *Unsupervised* mode, as in ELMER 1.0, for each probe-gene pair, the samples (all samples from both groups) are divided into two groups: the M group, which consist of the upper methylation quintile (the 20%of samples with the highest methylation at the enhancer probe), and the U group, which consists of the lowest methylation quintile (the 20% of samples with the lowest methylation). In the new *Supervised* mode, the U and M groups are defined strictly by sample group labels, and all samples in each group are used. The *Supervised* mode can greatly increase statistical power, as illustrated in Supplementary Figure 2.

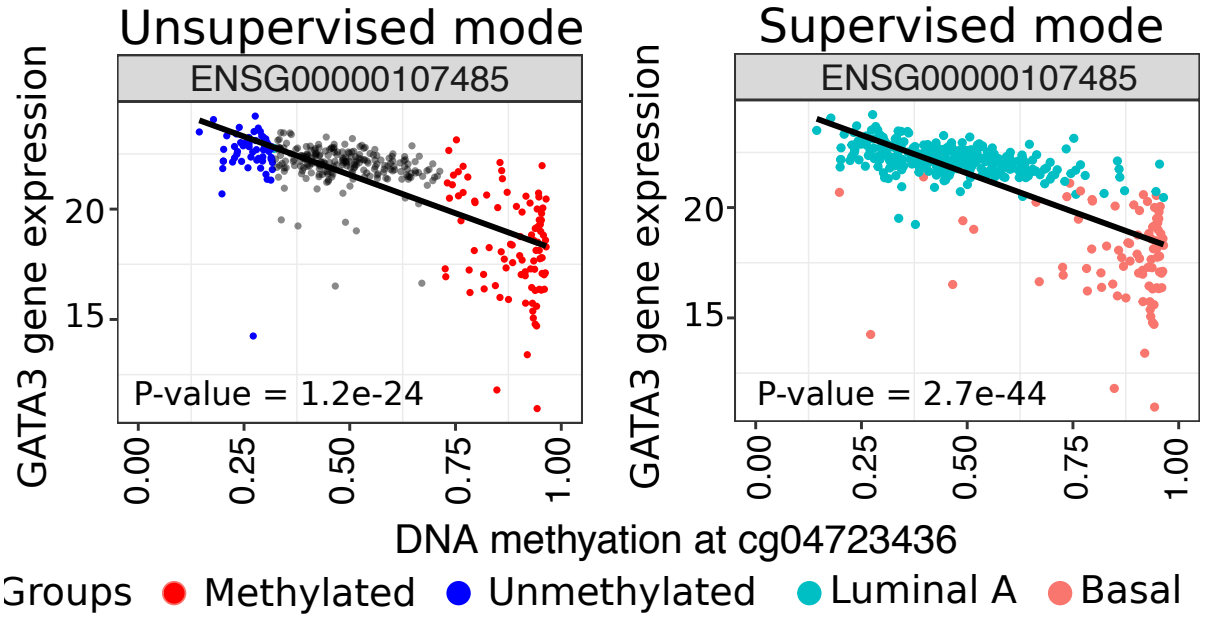
For each differentially methylated distal probe (DMC), the closest 10 upstream genes and the closest 10 downstream genes are tested for inverse correlation between methylation of the probe and expression of the gene (the number 10 can be changed using the *numFlankingGenes* parameter). To select these genes, the probe-gene distance is defined as the distance from the probe to the transcription start site specified by the ENSEMBL gene level annotations (Yates et al., 2015) accessed via the R/Bioconductor package biomaRt (Durinck et al., 2005, 2009). By choosing a constant number of genes to test for each probe, our goal is to avoid systematic false positives for probes in gene rich regions. This is especially important given the highly non-uniform gene density of mammalian genomes.

Thus, exactly 20 statistical tests were performed for each probe, as follows.

For each candidate probe-gene pair, the Mann-Whitney U test is used to test the null hypothesis that overall gene expression in group M is greater than or equal than that in group U. This non-parametric test was used in order to minimize the effects of expression outliers, which can occur across a very wide dynamic range. For each probe-gene pair tested, the raw p-value P_r is corrected for multiple hypothesis using a permutation approach as follows. The gene in the pair is held constant, and x random methylation probes are chosen to perform the same one-tailed U test, generating a set of x permutation p-values P_p. We chose the x random probes only from among those that were "distal" (farther than $2kb$ from an annotated transcription start site), in order to draw these null-model probes from the same set as the probe being tested (Sham and Purcell, 2014). An empirical p-value P_e value was calculated using the following formula (which introduces a pseudo-count of 1):

$$P_e = \frac{num(P_p \leq P_r) + 1}{x + 1} \tag{1}$$

Notice that in the *Supervised* mode, no additional filtering is necessary to ensure that the M and U group segregate by sample group labels. The two sample groups are segregated by definition, since these probes were selected for their differential methylation, with the same directionality, between the two groups.



Supplementary Fig 2. Supervised mode maximizes statistical power. Difference of groups U and M definition in *supervised* and *unsupervised* mode. A: *unsupervised* mode; when minSubgroupFrac argument is set to 40%, the methylated group is defined as the highest quintile and the unmethylated group as the lowest quintile; B: *supervised* mode; methylated and unmethylated group are defined as one of the known molecular subtypes. For example, the unmethylated group is represented by all the LumA samples while the methylated group is represented by all the Basal samples. The t-test p-value achieved for the Unsupervised mode is $1.2E-24$, while the Supervised mode is : $2.7E-44$.

Characterization of chromatin state context of enriched probes using StateHub

While ELMER version 1 was limited to searching within annotated enhancer elements, we have since found that this constraint was not necessary to achieve statistical power. Thus ELMER 2.0 by default searches *all* distal elements in the genome (distal elements are those greater than $\pm 2kb$ from a TSS; By changing ELMER default settings, it is possible to analyze TSS-proximal probes, either together with distal probes or separately. See ELMER Bioconductor documentation for details).

Because ELMER can now search essentially all probes on the array, it is important to understand the context of the probes that result from an ELMER analysis. Typically these are enhancer probes, but some regulatory changes may involve promoters, insulators, etc. We used the *StateHub* (Coetzee et al., 2017) and *FunciVar* Bioconductor packages to characterize enrichment of the various cell-type-specific chromatin states in the significant BRCA-hypomethylated probes (Supplementary Figures 3, 4). Importantly, the MCF-7 cell line, and ER-positive breast cancer cell line, is much more strongly enriched for all enhancer and promoter classes than other cell types. As more reference cell types become available, this analysis will be useful in characterizing tumor GRN changes that reflect particular cell types or co-opted developmental programs.

All methods are described here: `https://www.simoncoetzee.com/bioc2017.html` FunciVar by default calculates a likelihood based on the beta-binomial distribution, returning a 95% credible interval (optionally set by the "CI" argument) for the range of differences between the two populations of variants (i.e. foreground and background). Specifically it calculates a distribution of true enrichment (as probability of overlap) for both sets of variants in the genomic features based on the observed number of overlaps:

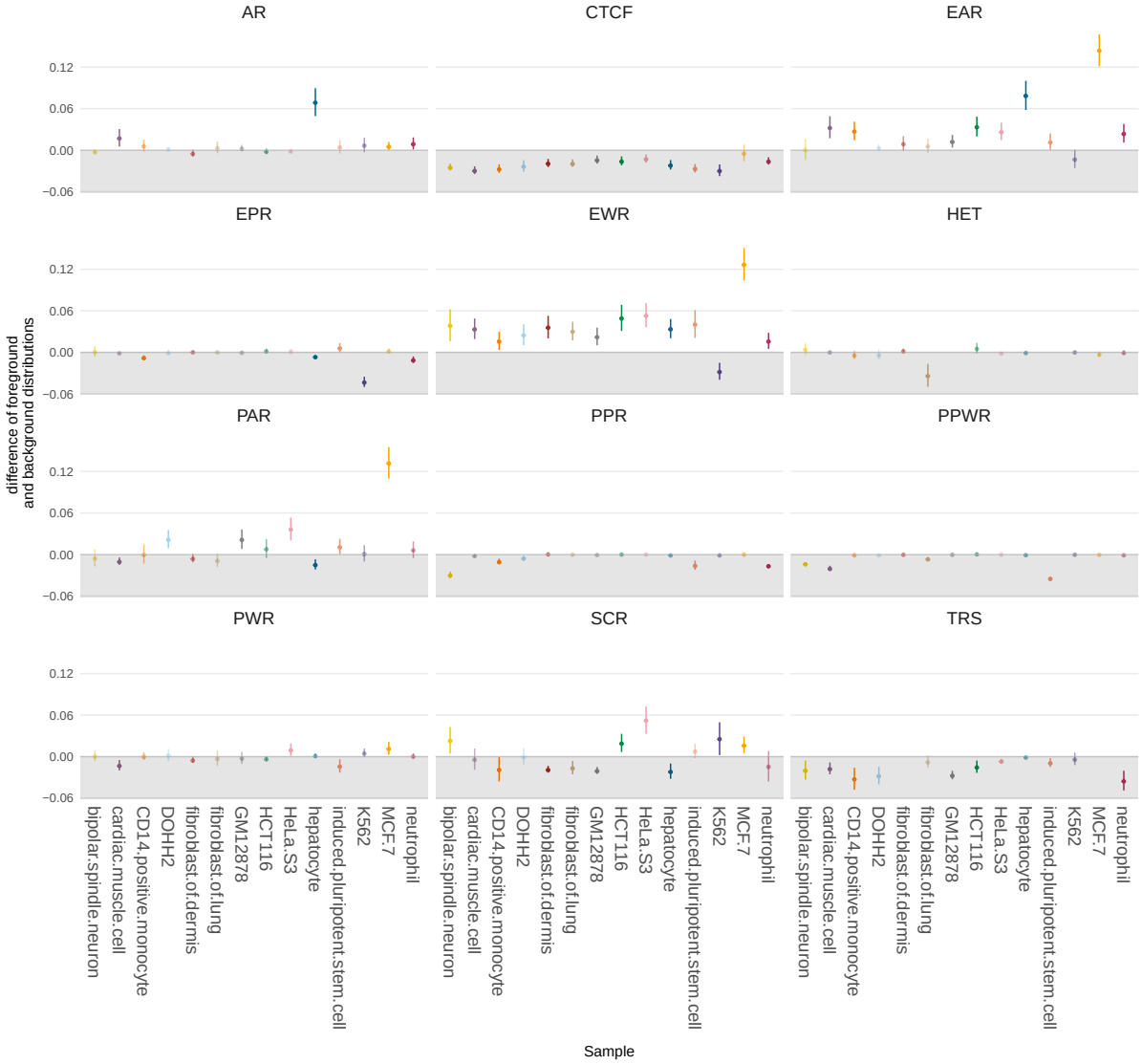
$$\theta_{fg} Beta(S_{fg} + \alpha, N_{fg} + \beta) \tag{2a}$$

$$\theta_{bg} Beta(S_{bg} + \alpha, N_{bg} + \beta) \tag{2b}$$

for S successes in N trials. FunciVar uses an uninformative Jeffreys prior c(a=0.5, b=0.5) to compare the two distributions directly by subtracting permuted samples to obtain the distribution of differences. The prior can be overridden in special cases.

☐	Track Name	Description	Project	Genome	Marks	Stat…	Get
☐	bipolar_spindle_neuron	bipolar spindle neuron	ENCODE	hg38	CTCF,H3K27ac,H3K36me3,H3K9me3,…	0.3	download
☐	cardiac_muscle_cell	cardiac muscle cell	ENCODE	hg38	CTCF,H3K4me1,H3K4me3,H3K27ac,H…	0.3	download
☐	cd14-positive_monocyte	cd14-positive monocyte	ENCODE	hg38	CTCF,H3K27ac,H3K27me3,H3K36me3…	0.3	download
☐	dohh2	dohh2	ENCODE	hg38	CTCF,H3K36me3,H3K27me3,H3K4me…	0.3	download
☐	fibroblast_of_dermis	fibroblast of dermis	ENCODE	hg38	CTCF,H3K4me3,H3K9ac,H3K9me3,H3…	0.3	download
☐	fibroblast_of_lung	fibroblast of lung	ENCODE	hg38	CTCF,H3K4me3,H3K4me3,H3K9ac,H3…	0.3	download
☐	gm12878	gm12878	ENCODE	hg38	CTCF,H3K4me3,H3K4me3,H3K36me3,…	0.3	download
☐	hct116	hct116	ENCODE	hg38	CTCF,H3K27ac,H3K4me3,H3K4me1,H…	0.3	download
☐	hela-s3	hela-s3	ENCODE	hg38	CTCF,H3K27ac,H3K9me3,H3K27me3,…	0.3	download
☐	hepatocyte	hepatocyte	ENCODE	hg38	CTCF,H3K27ac,H3K9me3,H3K36me3,…	0.3	download
☐	induced_pluripotent_stem_cell	induced pluripotent stem cell	ENCODE	hg38	CTCF,H3K4me3,H3K4me1,H3K27ac,H…	0.3	download
☐	k562	k562	ENCODE	hg38	CTCF,CTCF,CTCF,CTCF,H3K4me3,H3K…	0.3	download
☐	mcf-7	mcf-7	ENCODE	hg38	CTCF,CTCF,CTCF,CTCF,H3K27me3,H3…	0.3	download
☐	neutrophil	neutrophil	ENCODE	hg38	CTCF,H3K27me3,H3K4me1,H3K4me3,…	0.3	download

Supplementary Fig 3. Chromatin states were retrieved from http://statehub.org/ from the Default Model (Decision Matrix). Only states from encode samples having H3K27ac, H3K4me1, H3K4me3 and CTCF marks for hg38 were used.



Supplementary Fig 4. Enrichment of paired probes and chromatin states of encode cells. The plot shows enrichment for enhancer active region, weak enhancer and active promoter region for MCF-7 cell. Supplementary Figure 3 summarizes Chromatin states used in the enrichment analysis. Acronyms - AR: Active region, EAR: active enhancer, EWR: Weak Enhancer, EPR: poised enhancer, PAR: active promoter, PWR: Weak Promoter, PPR: poised promoter, PPWR: Weak Poised Promoter, CTCF: architectural complex, TRS: transcribed, HET: heterochromatin, SCR: Polycomb Repressed Silenced. Y-axis shows the probability difference in overlap for the foreground class vs. random probes (Confidence Interval based on beta-binomial distribution, see methods).

Motif enrichment analysis

In order to identify enriched motifs and potential upstream regulatory TFs, all probes with occurring in significant probe-gene pairs are combined for motif enrichment analysis. HOMER (Hypergeometric Optimization of Motif EnRichment) (Heinz et al., 2010) is used to find motif occurrences in a $\pm 250bp$ region around each probe, using HOCOMOCO (HOmo sapiens COmprehensive MOdel COllection) v11 (Kulakovskiy et al., 2016). Transcription factor (TF) binding models are available at `http://hocomoco.autosome.ru/downloads`. HOCOMOCO is the most comprehensive TFBS database and is consistently updated, marking an improvement over ELMER version 1.

For each probe set tested (i.e. the set of all probes occurring in significant probe-gene pairs), we quantify enrichments using Fisher's exact test (where a is the number of probes within the selected probe set that contains one or more motif occurrences; b is the number of probes within the selected probe set that do not contain a motif occurrence; c and d are the same counts within the entire array probe set drawn from the same set of distal-only probes using the same definition as the primary analysis) and multiple testing correction with the Benjamini-Hochberg procedure (Fisher, 1922).

A probe set was considered significantly enriched for a particular motif if the 95% confidence interval of the Odds Ratio was greater than 1.1 (specified by option *lower.OR*, 1.1 is default), the motif occurred at least 10 times (specified by option *min.incidence*, 10 is default) in the probe set and $FDR < 0.05$.

Identification of master regulator TFs

When a group of enhancers is coordinately altered in a specific sample subset, this is often the result of an altered upstream *master regulator* transcription factor in the gene regulatory network. *ELMER* identifies master regulator TFs corresponding to each of the TF binding motifs enriched from the previous analysis step. For each enriched motif, *ELMER* takes the mean DNA methylation of all distal probes (in significant probe-gene pairs) that contain that motif occurrence (within a $\pm 250bp$ region) and compares this mean DNA methylation to the expression of each gene annotated as a human TF.

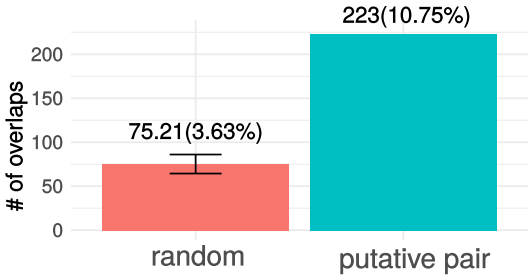
In the *Unsupervised* mode, a statistical test is performed for each motif-TF pair, as follows. All samples are divided into two groups: the M group, which consists of the 20% of samples with the highest average methylation at all motif-adjacent probes, and the U group, which consisted of the 20% of samples with the lowest methylation. This step is performed by the *get.TFs* function, which takes *minSubgroupFrac* as an input parameter, again with a default of 20%. For each candidate motif-TF pair, the Mann-Whitney U test is used to test the null hypothesis that overall gene expression in group M is greater or equal than that in group U. This non-parametric test was used in order to minimize the effects of expression outliers, which can occur across a very wide dynamic range. For each motif tested, this results in a raw p-value (P_r) for each of the human TFs.

The new *Supervised* mode uses the same approach as described for the identification of putative target gene(s) step. The U and M groups are one of the the label group of samples and the *minSubgroupFrac* parameter is set to 100% to use all samples from both groups in the statistical test. This also can result in greater statistical power when using the *Supervised* mode.

Comparing inferred results with MCF-7 ChIA-PET

As in our earlier paper, we compared CRM / gene pairs identified by Unsupervised analysis of TCGA Breast Cancer cases to chromatin loops derived from deep-sequenced ChIA-PET data from ER+ Breast Cancer MCF7 cells Li et al. (2012). ELMER pairs were enriched for ChIA-PET loops by roughly 3-fold over random pairs, consistent with our earlier results (Supplementary Fig. 5).

First, we identify the number of *ELMER* pairs overlapping the ChIA-PET loops, then we repeat using randomly generated pairs with properties similar to the *ELMER* pairs. For each true ELMER probe in a probe-gene pair, we randomly select a different probe from the complete set of distal probes. We then choose the nth nearest gene to the random probe, where n is the same as the adjacency of the true ELMER probe (i.e. if the true probe is linked to the second gene upstream, the random probe will also be linked to its second gene upstream). Thus, the random linkage set has both the same number of probes and the same number of linked genes as the true set. One hundred such random datasets were generated to arrive at a 95% CI ($\pm 1.96 * SD$). The result is shown in Supplementary Fig. 5. Of the 2118 putative pairs identified in breast cancer tumors, 223 ($\approx 10.75\%$) were also identified as loops in the MCF7 ChIA-PET data. This was a three-fold enrichment over randomized probe-gene pairs.

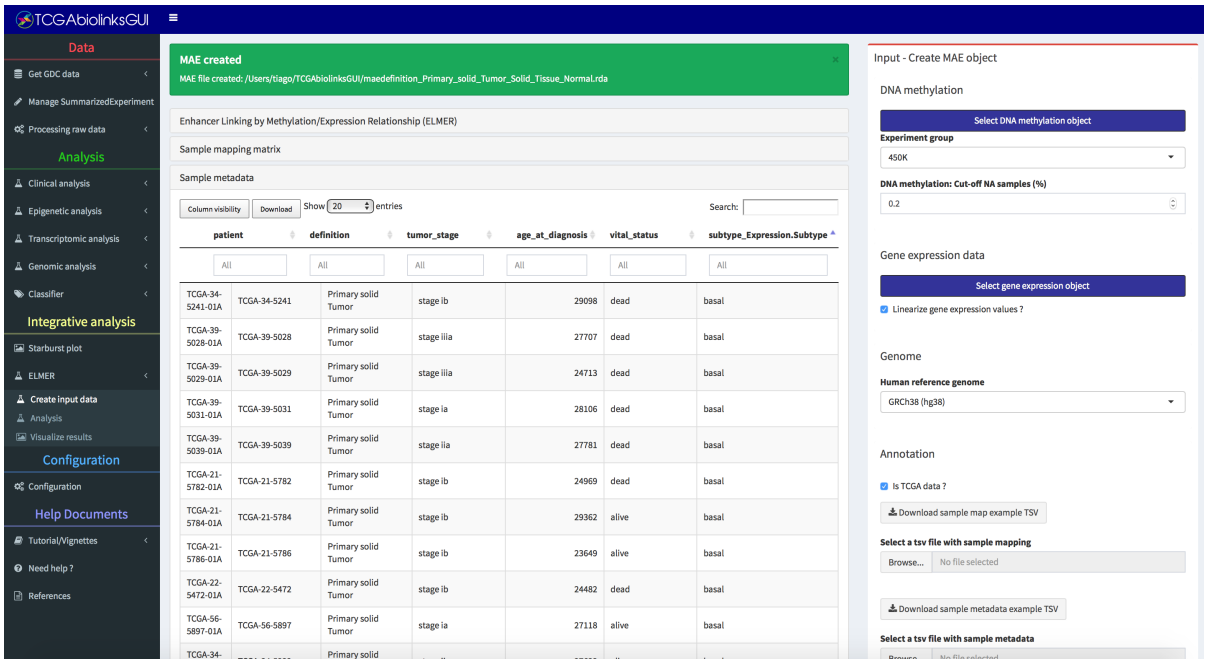


Supplementary Fig 5. The graph shows the comparison of the number of probe-gene pairs identified within MCF7 ChIA-PET data using the putative pairs from BRCA vs. random pairs.

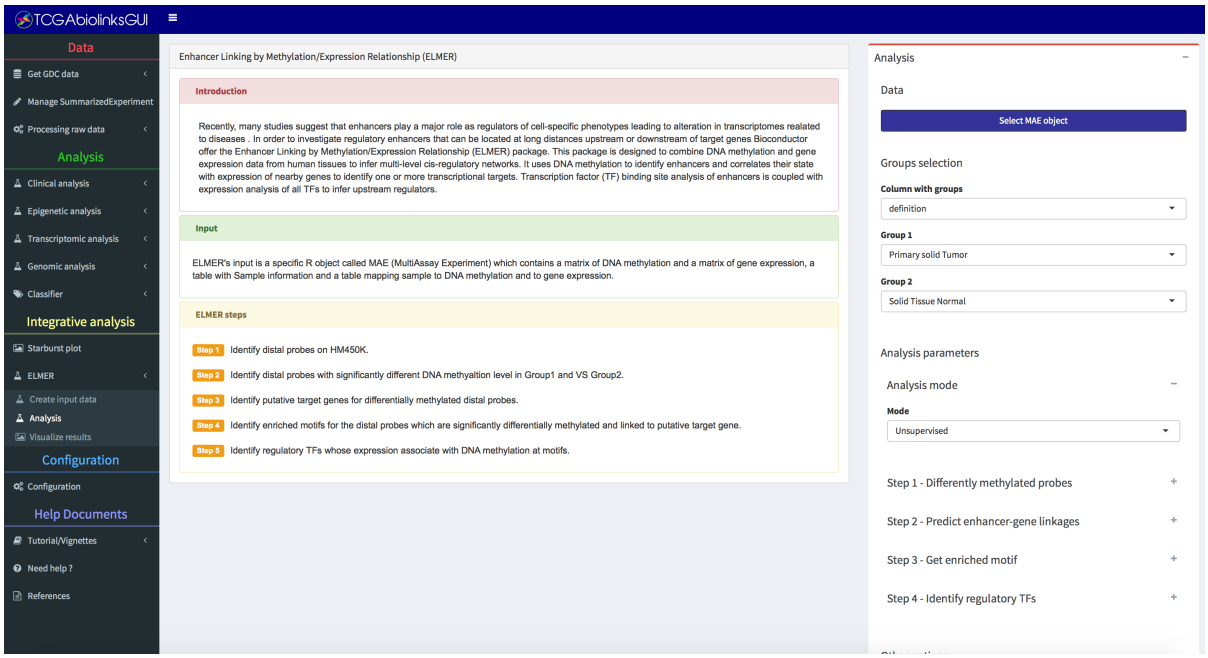
Graphical User Interface

To enable user access to the methodologies offered in ELMER and to give users the flexibility of point-and-click style analysis without the need to learn R, we have implemented a full graphical user interface (GUI) through the R/Bioconductor package TCGAbiolinksGUI (Silva et al., 2017) available at `http://bioconductor.org/packages/TCGAbiolinksGUI/`. This tool allows definition of sample groupings based on user-defined clinical attributes in the supported databanks, including TCGA and TARGET, and GDC. A tutorial detailing the steps needed to use the tools through the GUI is available at `https://bioinformaticsfmrp.github.io/Bioc2017.TCGAbiolinks.ELMER/index.html`

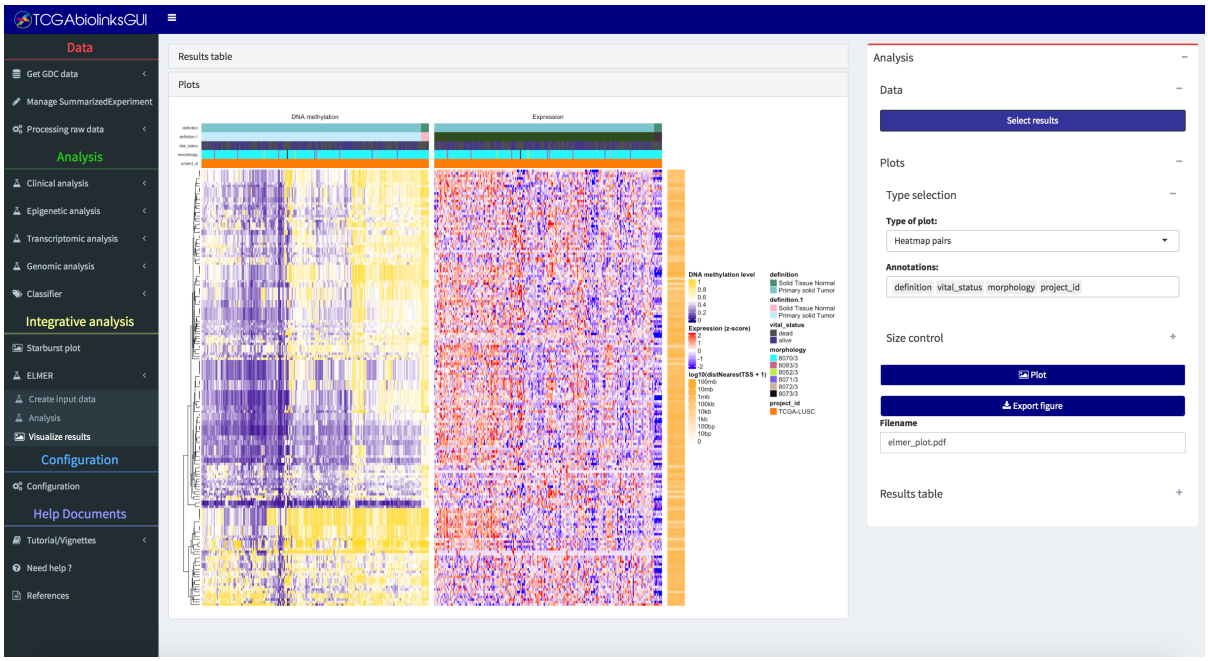
Supplementary Figures 6,7 and 8 shows the three ELMER menus in the R/Bioconductor TCGAbiolinksGUI package.



Supplementary Fig 6. ELMER graphical user interface in TCGAbiolinksGUI: MAE creation menu.



Supplementary Fig 7. ELMER graphical user interface in TCGAbiolinksGUI: analysis menu.

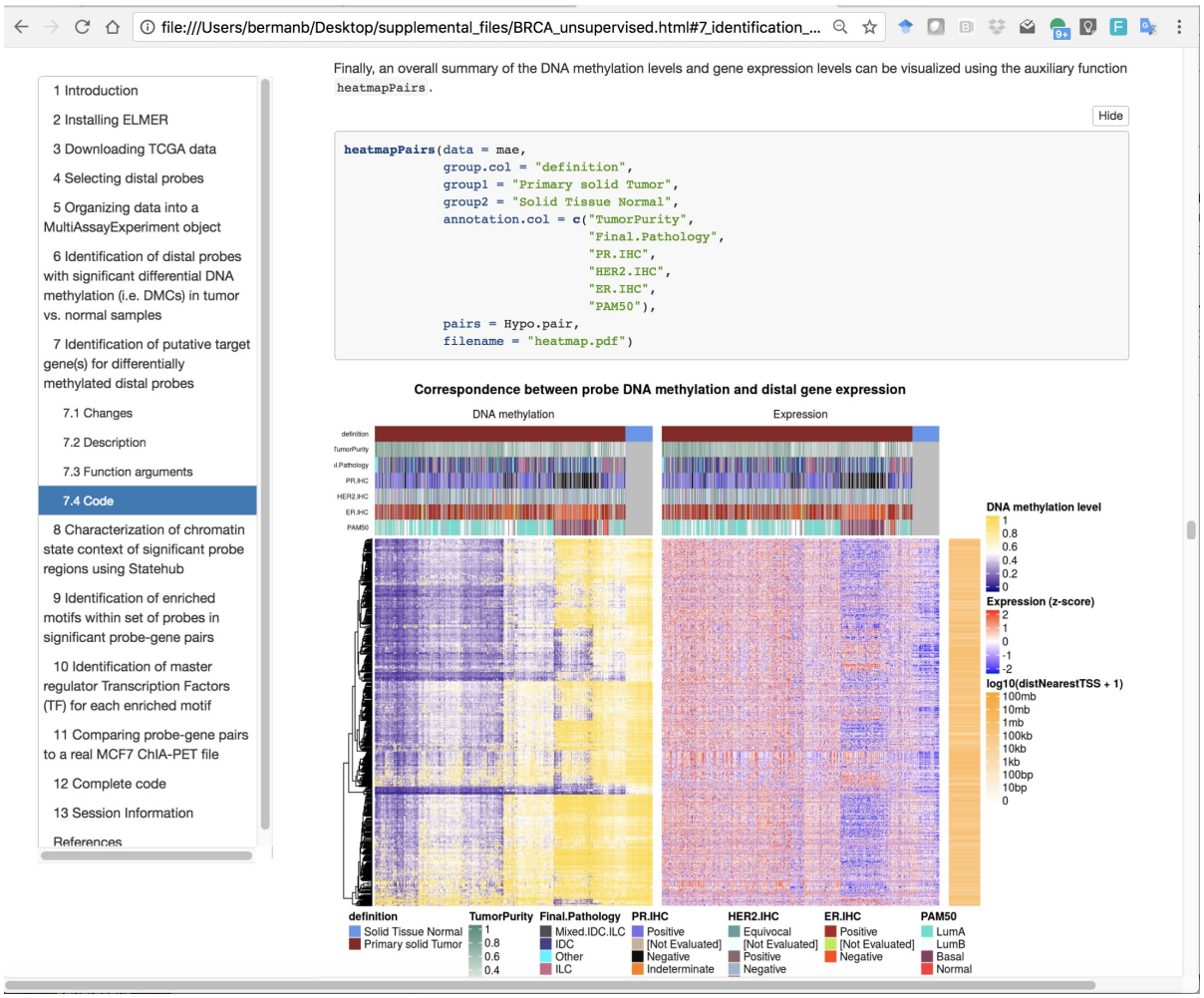


Supplementary Fig 8. ELMER graphical user interface in TCGAbiolinksGUI: visualization menu.

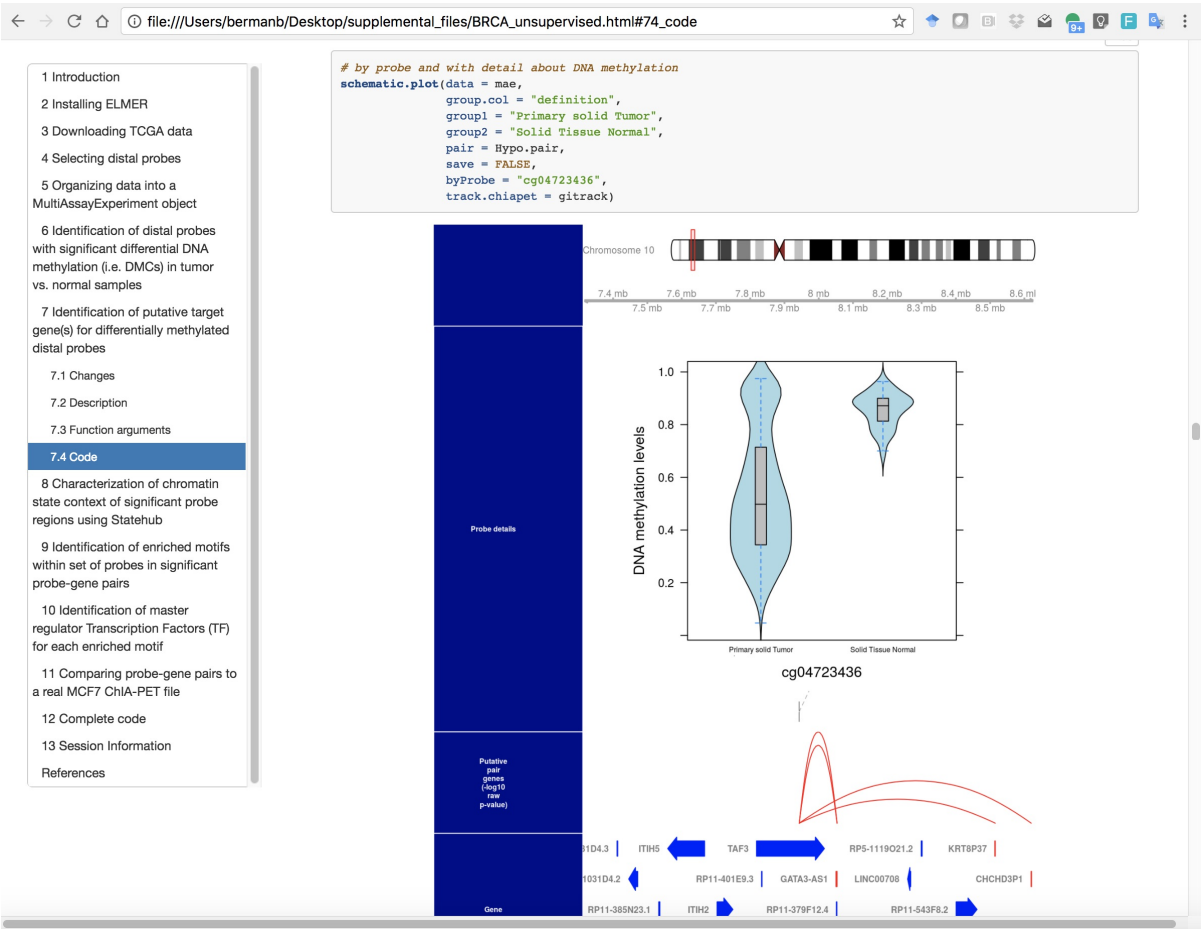
Interactive HTML output reports

While ELMER version 1 had functions to create individual output plots, we have completely revamped and added to the set of automatic output plots, which were used to generate all the Figures and Supplemental Figures in this article. We also now output a single HTML file which contains all output code, output tables, and plots for an individual ELMER run. This HTML file is indexed via a table of contents, and individual sections can be expanded and compressed to expose additional detail.

Supplementary Figs 9 and 10 shows the single interactive HTML file output for the Unsupervised analysis above. This and the HTML file for the Supervised analysis can both be downloaded at `https://github.com/tiagochst/ELMER_supplemental/raw/master/supplemental_files.zip`.



Supplementary Fig 9. Single HTML file output report example, showing generation of the comprehensive heatmap plot.

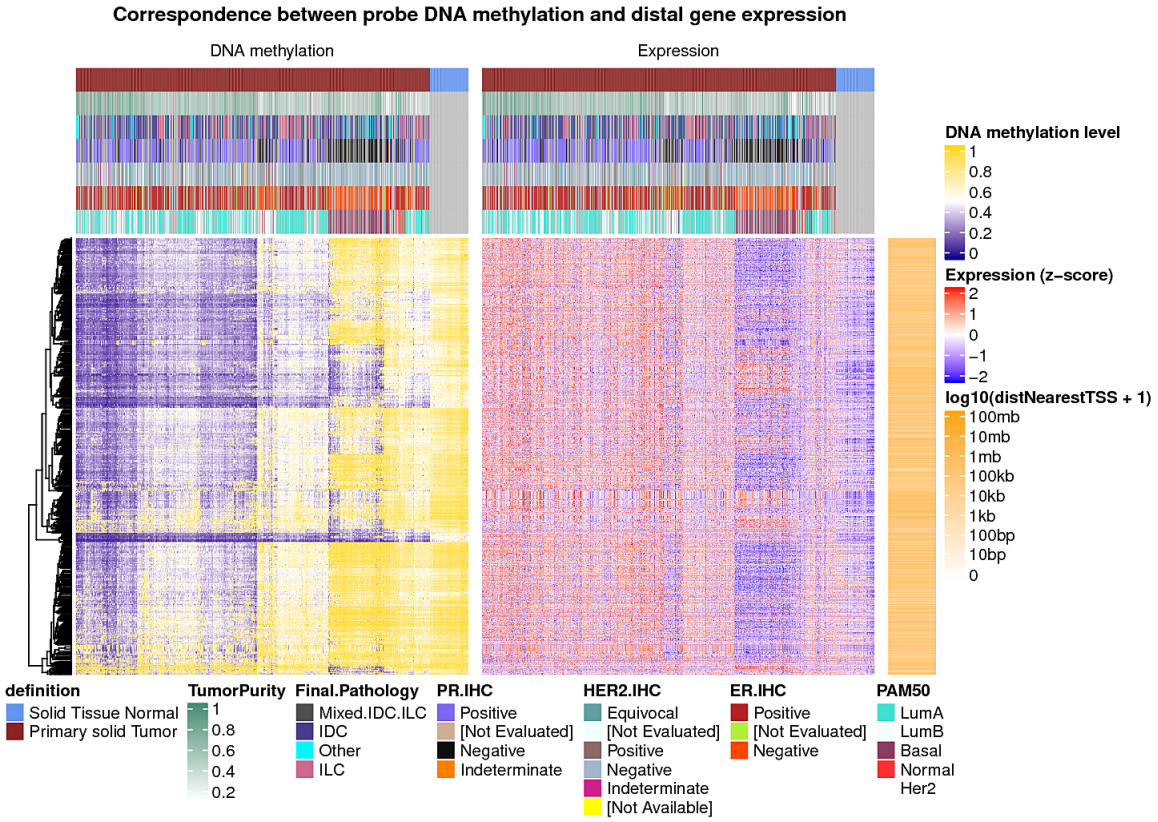


Supplementary Fig 10. Single HTML file output report example, showing generation of the genome browser plot.

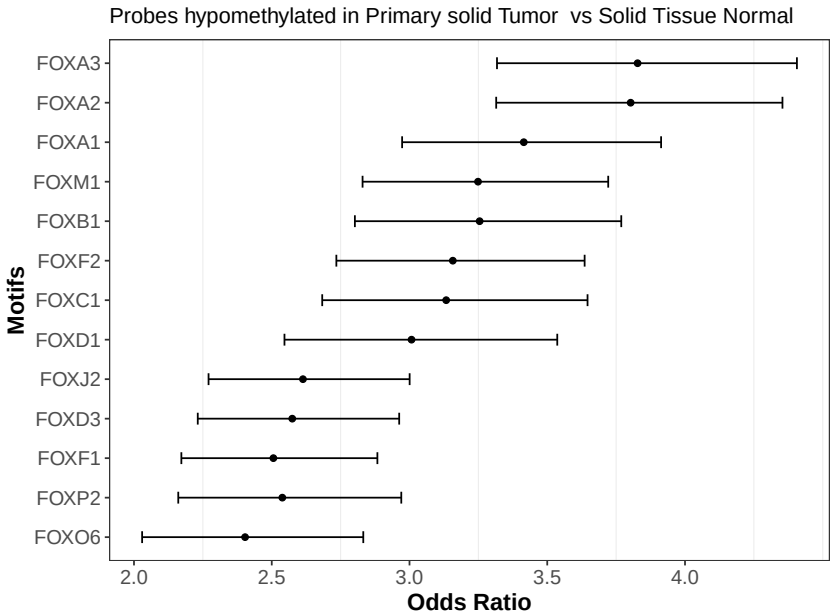
Use Case 1: Breast Invasive Carcinoma (Unsupervised mode)

We performed *ELMER* analysis comparing 778 Breast Invasive Carcinoma (Primary solid tumor) samples to 83 samples of normal tissue adjacent to the tumor In this use case we wanted to be able to detect non pre-determined molecular subtypes among the tumor samples, so the percentage of samples used to identify the differentially methylated probes in function *get.diff.meth* was set to 20% and the mode in function *get.pair* and in function *get.TFs* which was set to "unsupervised". In this mode we define the U (unmethylated) group as the samples with lowest quintile of DNA methylation levels and the M (methylated) group as the highest quintile.

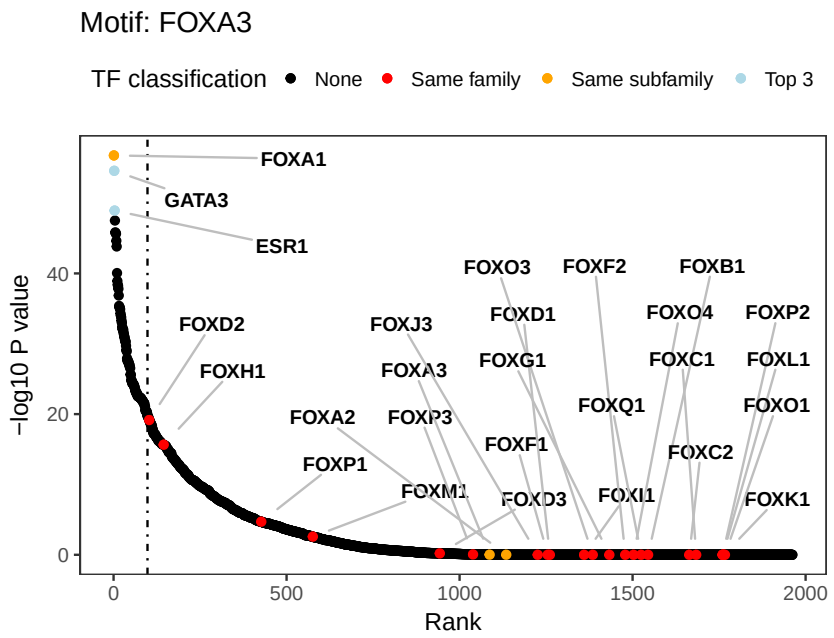
This analysis showed that the set of hypomethylated CpG probes (DMCs) in the tumors and linked to the expression of a nearby gene (Supplementary Figure 11) had an enrichment for TFBS motifs for FOX family transcription factors (*FOXA2*, *FOXA3*, *FOXA1*, etc.) (Supplementary Figure 12). For the most highly enriched motif *FOXA3*, the master regulator analysis identified *FOXA1* as the top candidate among all TFs in the human genome (Supplementary Figure 13), with the collaborating factors GATA3 and ESR1 as the next best candidates (Supplementary Figure 14). This illustrates the important point that *in vitro* defined motifs from public TFBS databases are not always bound by the same TF family member *in vivo*. This was the same result of the analysis in Yao et al., where we showed that ELMER identification of *FOXA1*, *GATA3*, and *ESR1* were driven specifically by the ER+ (luminal A and luminal B) tumors. However, our unsupervised analysis (both in Yao et al. and here) did not reveal Master Regulators for the other Breast Cancer molecular subtypes, such as Basal-like, HER2+, etc.



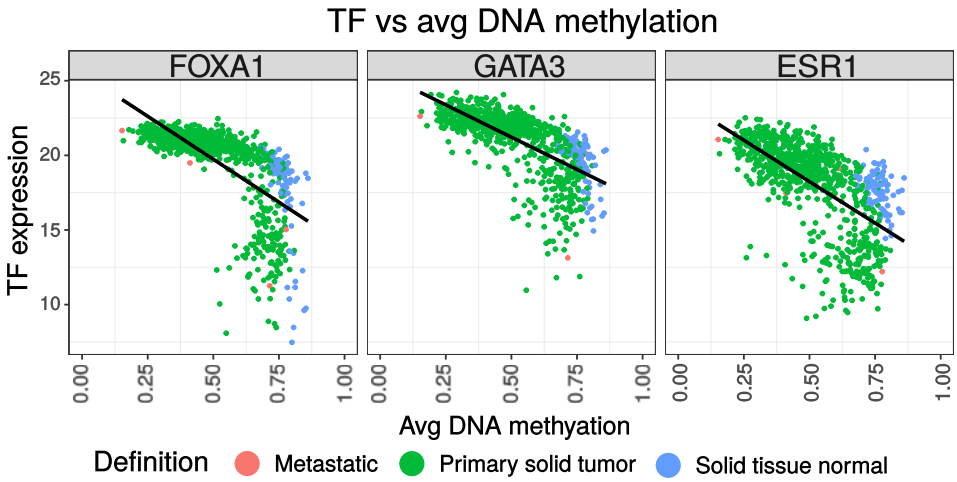
Supplementary Fig 11. The comprehensive heatmap view shows all probe / gene pairs identified by ELMER, clustered according to similarity. This plot is based on the Unsupervised analysis of Breast Invasive Carcinoma (Primary solid tumor) samples to 83 samples of normal tissue adjacent to the tumor (Solid Tissue Normal). The inverse correlation between methylation and expression can be observed.



Supplementary Fig 12. Motif enrichment plot shows the enrichment levels ($OR \geq 2.0$) for the most significant motifs based on the TCGA Breast Cancer Unsupervised analysis. A number of less significant motifs meet our default OR threshold of 1.1 ($lower.or = 1.1$), which can be browsed in our full Supplemental output report.



Supplementary Fig 13. TF ranking plot: For a given enriched motif, all human TF are ranked by the statistical $-log_{10}(P-value)$ assessing the anti-correlation level of candidate Master Regulator TF expression with average DNA methylation level for sites with the given motif. As a result, the most anti-correlated TFs will be ranked in the first positions. For example, the figure shows the TFs ranking for the (*FOXA3*) motif in which the TF *FOXA1* is ranked in the first position, meaning it is the most anti-correlated TF according to the statistical test. By default, the top 3 most anti-correlated TFs (*FOXA1*, *GATA3* and *ESR1*), and all TF classified by TFClass database in the same family (Forkhead box factors) and subfamily (FOXA) are highlighted with colors blue, red and orange, respectively. The anti-correlation data for the top three candidates (*FOXA1*, *GATA3*, and *ESR1*) are derived from the data shown in Supplementary Figure 14



Supplementary Fig 14. *FOXA1*, *GATA3* and *ESR1* were identified as the most significant Master Regulator candidates for the top motif (*FOXA3*). All FOX factors belonging to the same TFClass binding family are highlighted.

Use Case 2: BRCA molecular subtypes analysis (Supervised mode)

Several studies identified distinct molecular Breast Cancer molecular subtypes including luminal-like (Luminal A and Luminal B) subclasses, which are Estrogen receptor-positive (ER-positive), and the basal-like, ErbB2-positive and normal-like subclasses (ER-negative) (Perou et al., 2000; Sørlie et al., 2001; Yersal and Barutca, 2014).

We performed pairwise analysis comparing known molecular subtypes (Her2, Luminal A, Luminal B and Basal-like) using the TCGA BRCA dataset and classifications retrieved from Ciriello et al. (2015). Supplementary Table S2 shows the number of samples of each molecular subtype of breast cancer and Supplementary Table S3 summarizes the candidate MRs identified.

The unsupervised analysis of the same sample identified several Luminal type Master Regulators (MRs) such as FOXA1, GATA3, and ESR1. In order to identify MRs for the other subtypes, we created a table (Table S3) of candidate MRs identified by each pairwise ELMER run (complete results can be found in the supplemental HTML file described in the Supplementary Methods section).

Interestingly, several new MRs are identified for the Basal-like group, and these were mostly consistent in comparisons against Luminal and HER2+ subtypes. One group of MRs identified are the *SOX10* and *SOX9* TF signatures. For these signatures, the regulatory TF candidate identified are the *SOX9* (Sry-related HMG box-9) TF and *SOX11* (Sry-related HMG box-11) TF; this correlation between basal-like and *SOX11* was recently described by Shepherd et al. (2016) and *SOX9* was described by Gong et al.. Most interestingly, we found *KLF5* to be a consistently predicted MR for the Basal-like breast subtype. *KLF5* is a master pluripotency factor of embryonic stem cells, and has been associated with a number of different cancers. In breast cancer, its overexpression has been linked to aggressive, ER-negative and basal-like breast cancers (Ben-Porath et al., 2008).

Table S2. Number of samples of the molecular subtypes of breast cancer

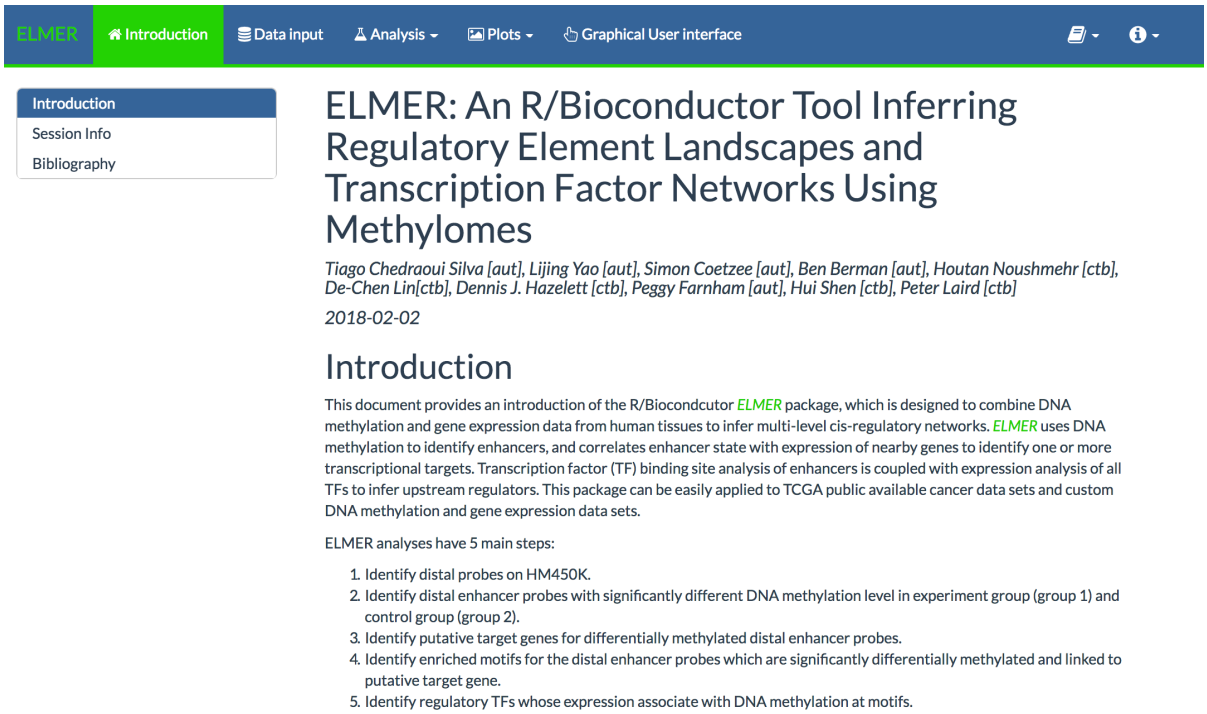
Molecular subtype	Number of samples
Basal	85
Her2	34
LumA	288
LumB	117
Normal-like	22

Table S3. Candidate regulatory TFs for each molecular subtype found in a pairwise comparison.

TF	LUMA (vs basal)	LUMB (vs basal)	LUMA (vs normal)	LumB (vs normal)	Basal (vs LumA)	Basal (vs LumB)	Basal (vs HER2)	HER2 (vs Basal)
AR	x	x	x					
BATF3						x		
BCL11A					x	x	x	
CBFB					x			
CEBPB					x	x	x	
CEBPG					x			
E2F3						x	x	
ELF5					x			
EMX1	x	x	x					
ESR1	x	x	x	x				
ELF5						x		
ETS2						x		
ETV6					x	x	x	
FOSL1						x		
FOXA1	x	x	x	x				x
FOXM1				x				
FOXD2		x						
FOXP1	x	x						x
GATA2	x							x
GATA3	x	x	x	x				x
GLI1	x	x						
HOMEZ		x						
HOXB1	x	x						
HOXB2	x	x						x
HOXB3								x
HOXB6								x
HOXC10								x
HOXC11								x
KLF5						x	x	
LMX1B	x	x	x					
MAZ			x					
MNX1								x
MSX2	x	x						
MYB	x			x				
MYBL1				x				
MYBL2				x	x			
NFATC4	x							
NFIB					x			
NFIL3					x	x		
NR2E3	x	x						
OVOL2			x					
PATZ1		x	x					
PBX1		x		x				
POU2F1			x	x				
PGR	x	x						
RARA	x	x	x					
RELB						x		
RORC	x							
RUNX3					x	x		
SOX8						x	x	
SOX9						x	x	
SOX11					x	x		
SPIB						x		
VEZF1		x						
ZBTB4	x							
ZNF281			x	x				
ZNF423	x							
ZNF467	x	x	x					
ZIC1					x	x	x	

Other improvements

To improve the software error handling we included in this new version: i) unit test which ensure our tool is working as expected and any modification in the code does insert bugs in the code, ii) TryCatch blocks to handle exceptions which will provide user with more information in case a exception is reached, iii) Continuous Integration (CI) services such as Travis `https://travis-ci.org/tiagochst/ELMER` and Appveyor (`https://ci.appveyor.com/project/tiagochst/elmer`) which not only ensures our tool is installable, free of bugs, passes unit tests and its documentation can be created after any code modification, but also reduces the time to identify possible platform specific problems. iv) the documentation has been revised and improved by changing the format of a PDF presentation to an HTML-navigable page (Supplementary Fig. 15), v) The data package documentation now has all the code required to easily create all auxiliary objects which was not available for ELMERv1.



Supplementary Fig 15. New ELMER documentation available at `http://bioconductor.org/packages/devel/bioc/vignettes/ELMER/inst/doc/index.html`

Time comparison

To compare the new version with the old one (ELMER v1 vs ELMER v2), we performed the unsupervised analysis of the TCGA breast cancer data set in an Ubuntu 16.04.3 LTS, 32Gb Intel Precision Tower 5810 Intel (R) Xeon (R) of RAM, CPU E5- 1650 v3 @ 3.50GHz, using 10 cores for code parallelization. The times for each one of the main functions are shown in Supplementary Table S4.

Some functions had an increase in the time due to changes either in the data or method. As the number of TF binding models used in this new version increased from 91 to 771 it was expected that the function *get.TFs* would increase the time to run, as more iterations will be performed. Also, the enriched.motif now performs and fisher test for each motif increasing the time to execute the function. Overall ELMER v2 decrease in 55% the time to run the analysis compared to ELMER v1. The code used to run ELMER v2 is the same provided in the HTML reports and the code to run ELMER v1 can be found in this gist (`https://gist.github.com/tiagochst/04c2c61b1f3f34f892cd0d0e12a81be6`).

Also, Supplementary Table S5 shows the time required to run each ELMER supervised analysis. Only runs that executed all functions were included, that means, if a analysis was not able to identify differential methylated probes it was excluded from the table. Although the larger the number of samples the longer the execution time, it is worth remembering that the unsupervised mode uses all samples in all the steps, while the supervised mode uses only a quintile of samples in each groups which will reduce its run time.

Table S4. Performance comparison between ELMER v1 vs ELMER v2. All values shown are in seconds.

Function	**Time elapsed ELMERv1**	**Time elapsed ELMERv2**
get.diff.meth	163 s	52 s
GetNearGenes	174 s	267 s
get.pair	29468 s	12790 s
get.enriched.motif	2 s	23 s
get.TFs	110 s	301 s
All functions	29917 s (≈ 8h18min)	13433 s (≈ 3h45min)

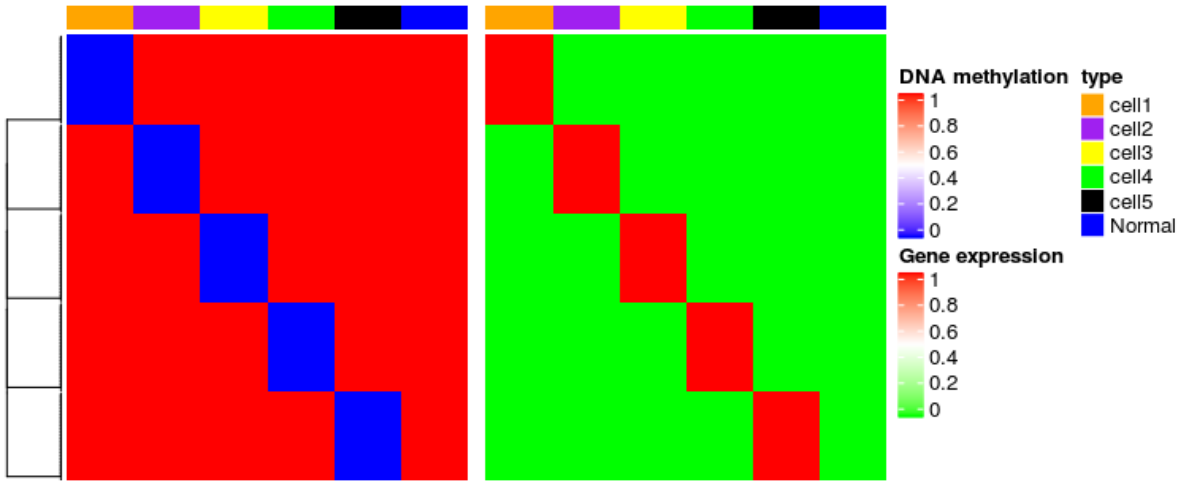
Table S5. ELMER functions runtime for each supervised analysis that executed all functions. All values shown are in seconds.

	Probes hypermethylated in group 1 vs group 2					**Probes hypomethylated in group 1 vs group 2**					
Function	**Basal vs Her2**	**Basal vs Normal**	**LumA vs Basal**	**LumB vs Basal**	**LumB vs Normal**	**Basal vs Her2**	**LumA vs Basal**	**LumA vs Normal**	**LumB vs Basal**	**LumB vs Normal**	**Normal vs Her2**
get.diff.meth	65	62	62	64	63	63	65	62	62	62	66
GetNearGenes	4	2	11	16	7	6	15	3	26	22	3
get.pair	211	113	2493	2618	891	611	3953	314	2901	1138	104
get.enriched.motif	10	10	13	15	11	11	15	10	15	12	10
get.TFs	20	49	75	85	98	50	190	58	231	163	100

Supervised mode vs *unsupervised* mode

It is assumed in the unsupervised mode that for at least one of the groups the samples belong to more than one molecular subtypes while in the supervised mode the assumption is that both groups have samples belonging to only one molecular subtype. For example, the unsupervised mode is more useful to perform analysis such as Normal vs Primary Solid Tumors, and the supervised LumA vs Basal-like.

To better explain the output of both modes, considering the hypothetical heatmap below (Supplementary Figure 16) and comparing the cell group (cell1, cell2, cell3) vs Normal, the unsupervised mode, which will use 20% of the samples, will identify all unmethylated regions (blue regions) and identify the anti-correlation with the gene expression (lost DNA methylation and the distal gene increased the expression). The next functions will identify the candidate master regulator TF that might be binding to those regions and activating the distal gene. But, that TF will be specific for only one molecular subtype and not all of them. Thus, it is expected that the unsupervised mode returns a general list of TF activated in group 1, but we are not able to differ in which molecular subtype it is activated. The supervised mode will solve that problem by comparing all samples from cell1 vs all Normal samples. Another drawback of the unsupervised mode is that regions specific to a certain molecular subtype might be not be identified if the number of samples is very low (i.e. 5% of all samples).



Supplementary Fig 16. Auxiliary heatmap for unsupervised mode vs Supervised explanation.

References

Aken, B. L. et al. (2016). The ensembl gene annotation system. *Database*, 2016:baw093.

Ben-Porath, I. et al. (2008). An embryonic stem cell–like gene expression signature in poorly differentiated aggressive human tumors. *Nature genetics*, 40(5):499–507.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the royal statistical society. Series B (Methodological)*, pages 289–300.

Ciriello, G. et al. (2015). Comprehensive molecular portraits of invasive lobular breast cancer. *Cell*, 163(2):506–519.

Coetzee, S. G. et al. (2017). Statehub-statepaintr: rapid and reproducible chromatin state evaluation for custom genome annotation. *bioRxiv*, page 127720.

Durinck, S. et al. (2005). Biomart and bioconductor: a powerful link between biological databases and microarray data analysis. *Bioinformatics*, 21(16):3439–3440.

Durinck, S. et al. (2009). Mapping identifiers for the integration of genomic datasets with the r/bioconductor package biomart. *Nature protocols*, 4(8):1184–1191.

Fisher, R. A. (1922). On the interpretation of χ 2 from contingency tables, and the calculation of p. *Journal of the Royal Statistical Society*, 85(1):87–94.

Gong, C. et al. Foxa1 repression is associated with loss of brca1 and increased promoter methylation and chromatin silencing in breast cancer. *Oncogene*, 34(39).

Heinz, S. et al. (2010). Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and b cell identities. *Molecular cell*, 38(4):576–589.

Huber, W. et al. (2015). Orchestrating high-throughput genomic analysis with bioconductor. *Nature methods*, 12(2):115–121.

Kulakovskiy, I. V. et al. (2016). Hocomoco: expansion and enhancement of the collection of transcription factor binding sites models. *Nucleic acids research*, 44(D1):D116–D125.

Li, G. et al. (2012). Extensive promoter-centered chromatin interactions provide a topological basis for transcription regulation. *Cell*, 148(1):84–98.

Perou, C. M., Sorlie, T., Eisen, M. B., Van De Rijn, M., et al. (2000). Molecular portraits of human breast tumours. *nature*, 406(6797):747.

Ramos, M. et al. (2017). Software for the integration of multi-omics experiments in bioconductor. *Cancer Research*, 77(21); e39-42.

Sham, P. C. and Purcell, S. M. (2014). Statistical power and significance testing in large-scale genetic studies. *Nature reviews. Genetics*, 15(5):335.

Shepherd, J. H. et al. (2016). The sox11 transcription factor is a critical regulator of basal-like breast cancer growth, invasion, and basal-like gene expression. *Oncotarget*.

Silva, T. C. et al. (2017). Tcgabiolinksgui: A graphical user interface to analyze gdc cancer molecular and clinical data. *bioRxiv*.

Sørlie, T. et al. (2001). Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. *Proceedings of the National Academy of Sciences*, 98(19):10869–10874.

Yao, L. et al. (2015). Inferring regulatory element landscapes and transcription factor networks from cancer methylomes. *Genome biology*, 16(1):105.

Yates, A. et al. (2015). Ensembl 2016. *Nucleic acids research*, page gkv1157.

Yersal, O. and Barutca, S. (2014). Biological subtypes of breast cancer: Prognostic and therapeutic implications. *World journal of clinical oncology*, 5(3):412.

Zhou, W. et al. (2016). Comprehensive characterization, annotation and innovative use of infinium dna methylation beadchip probes. *Nucleic Acids Research*, page gkw967.

Zhou, W. et al. (2017). Comprehensive characterization, annotation and innovative use of infinium dna methylation beadchip probes. *Nucleic Acids Research*, 45(4):e22.