VULCAN: Network Analysis of ER-α activation reveals reprogramming of GRHL2

Andrew N. Holding[1,3,*], Federico M. Giorgi[1,3], Amanda Donnelly[1], Amy E. Cullen[1], Luke A Selth[2], and Florian Markowetz[1]

[1] CRUK Cambridge Institute, University of Cambridge, Robinson Way, Cambridge, CB2 0RE
[2] Dame Roma Mitchell Cancer Research Laboratories and Freemasons Foundation Centre for Men's Health, Adelaide Medical School, The University of Adelaide, SA, Australia

[3] These authors contributed equally to this work.

*Corresponding Author

Abstract

VULCAN utilises network analysis of genome-wide DNA binding data to predict regulatory interactions of transcription factors. We benchmarked our approach against alternative methods and found improved performance in all cases. VULCAN analysis of estrogen receptor (ER) activation in breast cancer highlighted key components of the ER signalling axis and identified a novel interaction with GRHL2, validated by ChIP-seq and quantitative proteomics. Mechanistically, we show E2-responsive GRHL2 binding occurs concurrently with increases in eRNA transcription and GRHL2 negatively regulates transcription at these sites. These findings provide new insight into ER action in breast cancer and validate VULCAN as a powerful tool.

100/100

Keywords: Breast Cancer, Network Analysis, Dynamics, ER, Master Regulator, ChIP-seq, VULCAN, GRHL2, Squelching, P300

Introduction

Network Analysis of ChIP-seq to infer Transcription Factor (TF) activity

Usage of gene regulatory networks to analyse biological systems has witnessed an exponential increase in the last decade, due to the ease of obtaining genome-wide expression data [1–3]. Recently, the VIPER approach to interrogate these network models has been proposed to infer transcription factor activity using the expression of a collection of their putative targets, *i.e.* their regulon [4]. In the VIPER algorithm, gene-level *differential expression* signatures are obtained for either individual samples (relative to the mean of the dataset) or between groups of samples, and regulons are tested for enrichment.

Several tools exist to integrate ChIP-seq binding events to increase statistical power and to support the interpretation of data. These methods typically provide information in the context of biological pathways and established gene sets [5,6] or through motif analysis [7]. In contrast, we developed an extension of the VIPER algorithm, called "**V**irt**U**a**L** **C**hIP-seq **A**nalysis through **N**etworks" (VULCAN), to reveal and specifically analyse potential interactions of TFs in ChIP-seq experiments. Previously, the strategies employed by VIPER were limited to the analysis of transcription data. By developing VULCAN to overlay co-expression networks established from patient tumour data onto ChIP-seq data, we are able to provide candidate coregulators of the response to a given stimulus (**Figure 1**). Unlike previous method, the results of our analysis are therefore focused on discovering tissue or disease specific interactions in the context of these networks. Further, as VULCAN builds on master regulator analysis, the output from the pipeline provides the end user with information in terms of the activity of potentially interacting TFs, rather than in terms of pathways. The combination of disease specific context and TF activity information presents a significant step forward in providing

valuable information for the elucidation of on-chromatin interactions from ChIP-seq experiments over previous strategies.

The estrogen receptor as a model for systems biology

Breast cancer is the most common form of cancer in women in North America and Europe. The majority of breast cancers are associated with deregulated signalling by the Estrogen Receptor-alpha (ER), which drives tumour growth. ER is the key prognostic marker in the clinic and the target of the first lines of treatment in estrogen receptor positive (ER+) tumors. ER-targeting pharmacological strategies include SERMs (Selective Estrogen Receptor Modulators e.g. tamoxifen) and SERDs (estrogen receptor degraders, e.g. fulvestrant) and aromatase inhibitors that block the production of estrogens in the body [8]

On activation, ER binds to promoter and enhancer regions containing Estrogen-Response Elements (ERE) [9] to stimulate the transcription of hundreds of genes [10,11]. Gene expression is driven by both the recruitment of the basal transcription machinery to these loci and through longer range interactions [12]. Analysis of ER-target genes showed that many are proliferative in function and drive the growth of the tumour [13]

On the treatment of ER+ cells with estra-2-diol (E2), ER recruits several cofactors to form a complex on the chromatin. FOXA1 is of particular interest as the protein shares nearly 50% of its genomic binding sites with ER and has been shown to operate as a pioneer factor before ER activation [14]. It is through FOXA1 and other cofactors [15,16], e.g. SRC-1, that ER is able to recruit RNA Polymerase II at the gene promoter sites in order to initiate transcription [17].

Early studies have shown that, after stimulation with E2, ER binding to EREs can occur within minutes [18]. Maximum ER occupancy at promoters of target genes such as CTSD and

TFF1 is achieved after 45' in MCF7 cells. Roughly 90' after estradiol treatment, ER is partially released from the promoters of the target gene, only to reoccupy the site again at 135' and release them at 180', in a 45' phase cyclical manner [19]. The 90' occupancy phase has been shown to be independent of new protein synthesis, and is therefore thought to operate at the post-translational level. This cyclic, proteasome-mediated turnover of unliganded and liganded ER-α on responsive promoters is an integral feature of estrogen signaling [20,21].

Given the three features discussed above, ER is a prime target for systems biology. By applying VULCAN to ChIP-seq time-course data of ER activation, our study provides new temporal insights into ER cofactors on a genome-wide scale.

Results

The development of VULCAN to infer changes in cofactor activity from differential ChIP-seq analysis using network analysis led to novel insights into key coregulators of the ER binding process. An implementation of VULCAN in R is available on Bioconductor.org [https://bioconductor.org/packages/release/bioc/html/vulcan.html] and the scripts to replicate our analysis are available as a supplementary Rmarkdown file. Unless otherwise specified, all p-values were Bonferroni-corrected.

Regulatory network analysis to detect ER cofactors

VULCAN was developed in order to understand which TF or co-factors could be responsible for the temporal behavior of ER. To develop VULCAN, we combined co-expression networks, established from breast cancer tumour biopsy transcriptional data, with a modified version of

the VIPER algorithm [4] to enable master regulator analysis on differential binding signatures acquired from ChIP-seq analysis.

We generated co-expression networks using the most recent implementation of ARACNe [22] on the METABRIC dataset [23] and the independent TCGA dataset [24]. Briefly, ARACNe generates gene networks by estimating putative regulatory interactions between transcription factors and target genes using mutual information between gene expression profiles. As an example, **Figure 1C** shows an ARACNe-inferred targets of ESR1. The sets of targets of each TF (i.e. its regulon) were merged in a genome-wide transcriptional regulation network, as shown in a minimal diagram in **Figure 1D**.

Master regulator analysis [25] is an algorithm developed to identify transcription factors whose regulon is enriched within a list of differentially expressed genes. VULCAN extends the VIPER approach to use differential binding profiles rather than differential expression profiles. In this way, VULCAN can test the enrichment of TF regulons in ER occupancy signatures derived from ChIP-seq experiments (**Figure 1E**).

Comparison of VULCAN to existing methods

Comparing VULCAN's Mutual Information networks to Partial Correlation networks

VULCAN uses mutual information networks like VIPER [4]. To test the robustness of our approach to different underlying networks, we compared mutual information networks with partial correlation networks using different correlation thresholds. We generated several partial correlation networks from the TCGA Breast Cancer data using the same input as the ARACNe network used by VULCAN. We tested the overlap of every partial correlation network with the ARACNe network using the Jaccard Index (JI) criterion (**Suppl. Figure 33**). Finally, we show how the Jaccard Index between partial correlation networks and the ARACNe network is always

significantly higher than expected by selecting random network edges (**Suppl. Figure 34**). This confirms previous observations that partial correlation and mutual information networks are highly similar [26]

Comparing VULCAN with alternative methods for target enrichment analysis

We compared VULCAN's Gene Set Enrichment Analysis (GSEA) approach [27] with three independent methods previously applied to benchmark VIPER [4]. The first implemented a t-test based method, which takes the targets of a TF and integrates their p-value in a specific contrast. The method is similar to VIPER but involves a Fisher p-value integration step. The integrated test lacks stringency and results in nearly all regulons as significantly enriched (**Suppl. Figure 35**). Second, we implemented a fraction of targets method, defining for every TF the fraction of their targets that are also differentially bound. This alternative to VULCAN ignores the MI strength of interaction and the individual strengths of differential bindings, reducing the resolving power of the algorithm (**Suppl. Figure 36**). Finally, we compared to a Fisher's Exact Method, which assesses the overlap between networks and significant differential binding. This method is too stringent (as observed in the original VIPER paper) [4] and even without p-value correction there are no significant results, even at low stringency, demonstrating the low sensitivity of the using a Fisher's Exact Method method (**Suppl. Figure 37**). In summary, VULCAN outperformed all three alternative methods we tested (t-test based; fraction of targets method; and Fisher's Exact Method) in our dataset.

Comparing VULCAN with Online Tools (GREAT, ISMARA & ChIP-Enrich)

To further validate our method, we compared the output of our GSEA analysis with different versions of promoter-enrichment approaches implemented by GREAT [5], ISMARA [7] and

ChIP-Enrich [6]. The VULCAN analysis shows a significant overlap in terms of detected pathways with the GREAT method (**Suppl. Figure 38**). ChIP-Enrich identifies enrichment of a number of TFs also predicted by VULCAN, but it fails to identify ESR1 as the top transcription factor affected by our experiment (**Suppl. Figure 39**). ISMARA succeeds at identifying ESR1 using a motif-based analysis, but does not identify other candidate binding TFs (**Suppl. Figure 40**). In summary, VULCAN outperforms both ISMARA and ChIP-Enrich, and significantly overlaps with GREAT, but provides additional value through inference of TF factor activity.

VULCAN analysis of ER activation

Differential Binding Analysis

We performed four replicated ChIP-seq experiments for ER at three timepoints: 0, 45 and 90' after estradiol treatment **(Figure 2)** in the MCF7 breast cancer cell line**.** The binding profile of ER at each time point was then compared between time points using differential binding analysis. Differential binding analysis (**Figure 2B,C**) identified 18,900 statistically significant binding events at 45' ($p < 0.05$). We observed the previously reported reduction of ER binding at 90' [20] on a genome-wide level (17,896 significant binding events), but with a smaller amplitude than previous gene-specific assays [21]. A potential cause for the smaller amplitude we observed in ER cycling may relate to the fact that ChIP-seq is not inherently quantitative, and hence the typical normalisation strategies applied to ChIP-seq data are likely to suppress global changes [28,29]. We therefore validated the ER binding behaviour with ChIP-qPCR (**Figure 2A**) and observed the same reduction in amplitude at specific binding events as was predicted by ChIP-seq. Another potential reason for the difference in amplitude was we did not treat α-Amanitin prior to treatment with E2 [21] as this perturbation would further separate the experimental condition from clinical interpretation**.**

We performed motif enrichment analysis (HOMER software) on ER binding sites detected by differential binding analysis. This analysis confirmed a strong enrichment for a single element, ERE, bound at both 45 and 90', with a corrected p-value of 0.0029 (**Figure 3F**). When clustered according to peak intensity, samples cluster tightly in two groups: treated and untreated (**Suppl. Figures 2, 3 and 4**), but treatment at 45 and 90' is detectably different on a genome-wide scale, as highlighted by Principal Component Analysis (**Suppl. Figures 5 and 6**).

VULCAN groups genes by temporal dynamics of ER binding

We leveraged the information contained in mutual information networks to establish TF networks enriched in the differential binding patterns induced by estradiol. From our temporal analysis of ER binding, we established four classes of binding pattern: early responders, repressed transcription factors, late responders and candidate cyclic genes (**Figure 3**).

Using VULCAN, we defined TF network activity of occupied regulatory regions (**Figure 3A**) according to the binding of ER within their promoter and enhancer regions (limited to 10kb upstream of the Transcription Starting Site to ensure gene specificity). We define as early responders TFs whose network is upregulated at both 45 and 90' (**Figure 3B**); these genes include AR, SP1 and CITED1. TFs with opposite behaviour (namely TFs whose negative/repressed targets in the ARACNe model are occupied by ER), or "repressed TFs", include GLI4, MYCN and RAD21 (**Figure 3C**). Some TFs appear to have their targets bound at 45', but then unoccupied at 90'. This "updown" behaviour is consistent with the cyclic properties of certain components of the ER DNA-binding complex observed previously, and therefore we dubbed them "candidate cyclic TFs" (**Figure 3D**). We also define a "late responders" category, defined as TFs active at 90 but not at 45'. While this category exists, and notably contains both

ESR1 and the known ESR1 interactor GATA3, it is just below the significance threshold at 45' (**Figure 3E**).

Validating VULCAN results on independent data

We repeated our TF network activity analysis of ER activation (Figure 3A-E) on an independent data from TCGA (**Suppl. Figures 16–20**). We again found enrichment for ESR1 as a later responder. GATA3 was detected as an early responder, likely as a result of the METABRIC result being only slightly below the significance threshold at 45'. Notably, PGR was shifted from an early responder to a late responder.

To ensure the robustness of results, we performed a joint analysis of data obtained from both networks. At 45' (**Figure 4A**) and 90' (**Figure 4B**) we identified robust candidates, specifically the ESR1, GATA3 and RARA networks were consistently activated by ER in both time points. On the other hand, some factors, including HSF1 and GRHL2, were significantly repressed in the joint analysis.

As a negative control, we used a different context ARACNe network, derived from the TCGA AML dataset. This network shows globally weaker enrichment scores and a weak positive correlation with the results obtained through breast cancer regulatory models (**Suppl. Figure 22**).

Pathway analysis of regulatory region binding

We performed a Gene Set Enrichment Analysis [27] and an associated Rank Enrichment Analysis [4] using the differential binding at gene regulatory regions (with time 0 as reference). Individual differential binding signatures for GSEA were calculated using a negative binomial test implemented by DiffBind [30]. The collective contribution of differentially bound sites

highlights several ER-related pathways [31–33] (**Suppl. Figure 23**) in both the GSEA and aREA analyses. The strongest upregulated pathway in both time points (**Table S2 and S5**) was derived via RNA-seq in an MCF7 study using estradiol treatment [32], confirming the reproducibility of our dataset.

Validating VULCAN results by quantitative proteomics

We tested the performance of VULCAN against a complementary experimental approach called RIME [34] combined with TMT [35] (qPLEX-RIME), which aims at identifying interactors of ER within the ER-chromatin complex. We generated ER qPLEX-RIME data from MCF7 cells treated with estradiol at both 45 and 90' and compared this with the VULCAN dataset, with the aim of identifying TFs upstream of the observed differential binding (**Suppl. Figure 31**). We found known ESR1 interactors with both methods, namely HDAC1, NCOA3, GATA3 and RARA with positive Network Enrichment Score (NES) [4], implying the TF's regulon is over-represented within the differentially bound genes. Conversely, GRHL2 was discovered with a negative NES. A negative NES implies the regulon is either significantly depleted within the set of differentially bound genes or that TF is established by ARACNe as having a negative correlation to the genes regulated at these differentially bound sites.

The GRHL2 Transcription Factor

In our analysis of ER dynamics, the GRHL2 transcription factor stood out. In both the METABRIC and TGCA networks, GRHL2 was significantly repressed; yet in our proteomics analysis, the protein was significantly increased. GRHL2 is a transcription factor that is important for maintaining epithelial lineage specificity in multiple tissues [36,37]. It has previously been predicted to exist in ER-associated enhancer protein complexes [38], but its

function in the ER signalling axis is unknown. Therefore, we set out to experimentally validate GRHL2 as an ESR1 cofactor, possibly with repression properties for the ER complex.

Our analysis shows that the genes occupied by the ER complex do not form part of the GRHL2 regulon, using both TCGA-derived and METABRIC-derived regulatory models. Our analysis highlights the small overlap between the ESR1 (Estrogen Receptor) and GRHL2 networks (**Suppl. Figure 28**), hinting at complementary signals not dependent on global network overlaps. In fact, there is only a weak, positive correlation between ESR1 and GRHL2 expression in the TCGA breast cancer dataset (**Suppl. Figure 29**) and also in the METABRIC breast cancer dataset (**Suppl. Figure 30**). Furthermore, GRHL2 has a visibly lower variance than ESR1: it does not change significantly in different PAM50 subtypes, although it is lower in normal compared to malignant tissue.

The low overlap between networks and the low correlation in expression profiles could be explained by the fact that GRHL2's role may not be related to its function as a transcription factor; rather, it may be controlled by other mechanisms like phosphorylation, subcellular localisation or on-chromatin interactions.

GRHL2 activity before and after stimulation with E2

Differential ChIP-seq analysis of GRHL2 binding between 0 and 45' indicated that GRHL2 binding is altered on treatment with E2 **(Figure 5A)**. VULCAN analysis of the GRHL2 differential binding showed a consistent Network Enrichment Score for both the TCGA- and METABRIC-derived networks for ER, but not for FOXA1 or GRHL2 **(Figure 5B)**. Individual analysis of peaks show that typically ER promoter sites, e.g. RARa, were not the target of this redistribution of GRHL2, as these sites were occupied by GRHL2 before E2 stimulation. Motif

analysis of the sites within increased GRHL2 occupancy showed enrichment for the full ERE (p-value = 1 x 10^{-86}) and the GRHL2 binding motif (p-value = 1 x 10^{-31}).

qPLEX-RIME analysis of GRHL2 interactions showed high coverage of the bait protein (>59%) and, in both the estrogen-free and estrogenic conditions, high levels of transcription-related protein interactors including HDAC1 (p-value = 6.4 x 10^{-9}), TIF1A (p-value = 6.4 x 10^{-9}), PRMT (p-value = 6.4 x 10^{-9}) and GTF3C2 (p-value = 4.6 x 10^{-9}). P-values given for estrogenic conditions and estrogen-free conditions were comparable. The only protein differentially bound to GRHL2 in estrogen-free versus estrogenic conditions was ER.

The specific nature of the detected ER-GRHL2 interaction suggests that ER does not regulate or interact with GRHL2 through the recruitment of alternative cofactors. This implies instead that ER recruits GRHL2 to certain genomic loci for a specific function. We therefore undertook a comparison of GRHL2 binding with public datasets. **(Figure 5C).** Our analysis showed that GRHL2 sites that responded to E2 were enriched for ER binding sites (in agreement with our qPLEX-RIME data) and FOXA1 (compatible with either an ER interaction or the previously reported interaction with MLL3 [38]). To establish, therefore, if the reprogramming of GRHL2 was primarily related to a transcriptional function or the previously described interaction with MLL3, we overlapped our GRHL2 data with that of published H3K4me1/3 [38] and P300 [39] cistromes. While H3K4me occupancy was consistent between conditions, we found P300 binding to be enriched at the E2 responsive GRHL2 sites. A more detailed analysis of the GRHL2 overlap with P300 sites showed the greatest co-occupancy of GRHL2/P300 sites was when both TFs were stimulated by E2 **(Figure 5D).** Moreover**,** overlap of GRHL2 peaks with ER ChIA-PET data [ENCSR000BZZ] showed that the GRHL2 responsive sites were enriched at enhancers over promoters **(Figure 5E).**

These findings suggested that the GRHL2-ER interaction was involved in transcription at ER enhancer sites. To explore this concept further, we investigated the transcription of enhancer RNAs at these sites using publicly available GRO-seq data [40] [GSE43836] **(Figure 5F).** At E2 responsive sites**,** eRNA transcription was strongly increased by E2 stimulation; by contrast, eRNA transcription was largely independent of E2 stimulation when the entire GRHL2 cistrome was considered. Analysis of a second GRO-seq dataset, GSE45822, corroborated these results **(Suppl. Figure S92)**.

To further explore how GRHL2 regulates ER enhancers, we measured eRNA expression at the GREB1, TFF1 and XBP1 enhancers after over-expression of GRHL2. At TFF1 and XBP1, increased GRHL2 resulted in reduced eRNA transcription **(Figure 6)** ($p < 0.05$, paired-sample, t-test). Conversely, eRNA production at the TFF1, XBP1 and GREB1 enhancers was moderately increased after GRHL2 knockdown **(Suppl. Figure 106).** Combining data from all three sites established the effect as significant by paired-sample rank test ($p = 0.04$, one-tailed paired-sample, wilcoxon test). Collectively, these data demonstrate that GRHL2 constrains specific ER enhancers.

Analysis of mRNA expression levels of TFF1, XBP1 and GREB1 by RT-qPCR revealed that overexpression of GRHL2 also led to a significant reduction in TFF1 and XBP1 ($p < 0.05$, paired-sample, t-test), while knockdown of GRHL2 led to a significant increase in GREB1 transcription ($p < 0.05$, paired-sample, t-test) **(Suppl. Figure 107)**.

Discussion

VirtUaL ChIP-seq Analysis through Networks – VULCAN

The VULCAN method is valuable for the discovery of transcription factors which have a role in the regulation of protein complexes that would otherwise remain hidden. The challenge of highlighting cofactors from a ChIP-seq experiment lays in the infeasibility of reliable proteomic characterisation of DNA-bound complexes at specific regions. On the other hand, while RNA-seq is arguably the most efficient technique to obtain genome-wide quantitative measurements, any transcriptomic approach cannot provide a full picture of cellular responses for stimuli that are provided on a shorter timescale than mRNA synthesis speed, such as the estradiol administration described in our study. VULCAN, by combining RNA-seq derived networks and ChIP-seq cistrome data, aims at overcoming limitations of both. Most notably, our method can work in scenarios where candidate cofactors do not have a well characterised binding site or do not even bind DNA directly.

By developing VULCAN, we have been able to rediscover known cofactors of the estradiol-responsive ER complex and predict and experimentally validate a novel protein-protein interaction.

Reprogramming of GRHL2 by ER

In the 4T1 tumour model, GRHL2 was found to be significantly downregulated in cells that had undergone EMT [36]. The same study showed that knockdown of GRHL2 in MCF10A – an ER-negative cell line – lead to loss of epithelial morphology. Overall, this suggested that the

GRHL2 transcription factor plays an essential role in maintaining the epithelial phenotype of breast cells. Similar results were observed with the MDA-MB-231 model, where expression of GRHL2 resulted in reversal of EMT [37]. This result has been recapitulated in hepatocytes, where GRHL2 was found to suppress EMT by inhibiting P300 [41]

Survival data for ER+ breast cancer (KMplotter, use gene expression, p=0.001) and ER- (KMplotter, use gene expression, p=0.035) shows that high GRHL2 has a negative impact on survival time in both contexts. The ability to suppress EMT has also been noted in prostate cancer, another cancer driven by a steroid hormone receptor (AR), and the genes regulated by GRHL2 are linked to disease progression [42].

Combined, these earlier data indicate a significant role for GRHL2 in the progression of breast cancer, but its role in the ER signalling axis has, until now, been unknown. Here, we show that GRHL2 constrains activity at a subset of ER enhancers. Overexpression of GRHL2 resulted in a significant decrease in eRNA production at the TFF1 and XBP1 enhancers and, in agreement with previous studies that correlate eRNA transcription with gene expression [43–45], we found the measured eRNA decrease was concurrent with a significant downregulation in the expression of the corresponding gene.

These results are consistent with previous findings that GRHL2 inhibits P300 [41] and, while the ER complex results in the activation of eRNA transcription at these sites, that GRHL2 plays a role in fine-tuning or modulating this process.

In breast cancer, GRHL2 has previously been shown to directly interact with FOXA1, which may contribute to tethering of the histone methyltransferase MLL3 and, consequently, epigenetic marks at GRHL2/FOXA1 binding sites [38]. Our analysis, however, showed no particular enrichment for H3K4me1/3 marks at E2 responsive GRHL2 sites compared to other GRHL2 binding sites and our proteomic analysis of interactors showed a strong association with

proteins related to transcription. Therefore, while GRHL2 ChIP-seq analysis shows that GRHL2 is already bound to a proportion of FOXA1 sites before treatment of cells with E2, we still observed an increase in GRHL2 binding at 45'. We proposed that these ER responsive sites are related to a role of GRHL2 in a transcriptional process independent of its interaction with MLL3. This was supported by evidence of a significant overlap with binding of the coactivator P300 and a pronounced increase in eRNA transcription on activation at E2 responsive GRHL2 sites.

Conclusion

VULCAN is built on state-of-the-art network analysis tools previously applied to RNA-seq data. By adapting network-based strategies to ChIP-seq data, we have been able to reveal novel information regarding the regulation of breast cancer in a model system.

The VULCAN algorithm can be applied generally to ChIP-seq for the identification of new key regulator interactions. Our method provides a novel approach to investigate chromatin occupancy of cofactors that are too transient or for which no reliable antibody is available for direct ChIP-seq analysis.

VULCAN enabled us to identify the reprogramming of GRHL2 by the ER on stimulation with E2. Further analysis showed the process to be unrelated to the previously reported interaction with FOXA1 and MLL3 [38]. Our conclusion, therefore, is that GRHL2 has a second, previously undescribed, role: negatively regulating levels of transcription at estrogen responsive enhancers **(Figure 7)**. Given the central role of the ER in breast cancer development and GRHL2's own ability to regulate EMT, the discovery that ER recruits GRHL2, leading to the

constraint of eRNA transcription, is an important step in enhancing our understanding of breast cancer and tumourigenesis.

Methods

Sample preparation

MCF7 cells were obtained from the CRUK Cambridge Institute collection, authenticated by STR genotyping and confirmed free of mycoplasma. All cells were maintained at 37 ºC, 5% CO_2. For each individual ChIP pull-down, we cultured 8 x 10^7 MCF7 cells (ATCC) across four 15 cm diameter plates in DMEM with 10% FBS, Glutamine and Penicillin/Streptomycin (Glibco). Five days before the experiment, the cells were washed with phosphate buffered saline (PBS) and the media was replaced with clear DMEM supplemented with charcoal treated serum. The media was refreshed every 24 hours, which halted the growth of the cells and ensured that majority ER within the cell was not active. On day 5, the cells were treated with estradiol (100 nM). At the appropriate time point, the cells were washed with ice cold PBS twice and then fixed by incubating with 10mL per plate of 1% formaldehyde in unsupplemented clear media for 10 minutes. The reaction was stopped by the addition of 1.5mL of 2.5 M glycine and the plates were washed twice with ice cold PBS. Each plate was then scraped in 1 mL of PBS with protease inhibitors (PI) into a 1.5 mL microcentrifuge tube. The cells were centrifuged at 8000 rpm for 3 minutes at 4 ºC and the supernatant removed. The process was repeated for a second wash in 1 mL PBS+PI and the PBS removed before storing at -80 ºC.

ChIP-Seq

Frozen samples were processed using established ChIP protocols [46] to obtain DNA fragments of ~300 bp in length. The libraries were prepared from the purified DNA using a Thruplex DNA-seq kit (Rubicon Genomics) and sequenced on the Illumina HiSeq Platform. Sequencing data is available from Gene Expression Omnibus, accession GSE109820.

Differential binding analysis

Sequencing data was aligned using BWA[47] to the human genome (hg19). Reads from within the DAC Blacklisted Regions was removed before peak calling with MACS 2.1 [48] on default parameters. The aligned reads and associated peak files were then analyzed using DiffBind [30] to identify significant changes in ER binding.

Gene Set Enrichment Analysis (GSEA)

Gene Set Enrichment Analysis (GSEA) was performed as described by *Subramanian et al*. [49] using the curated pathway collection (C2) from MSIGDB v 5.0 with 1000 set permutations for each pathway investigated, followed by Benjamini Hochberg P-value correction.

Motif analysis

Motif analysis of the binding regions was undertaken with Homer v4.4 [50] using default parameters. Motif logo rendering was performed using Weblogo v2.8.2 [51]

VULCAN analysis

We reconstructed a regulatory gene network using ARACNe-AP as described by Alvarez [26]. RNA-Seq breast cancer data was downloaded from TCGA on January 2015 and VST-Normalized as described by Anders and Huber [52]. The ARACNe transcriptional regulation network was imported into R using the *viper* BioConductor package and it was interrogated using the differential binding profiles from our ChIP-Seq experiment as signatures, 45' vs control and 90' vs control. The peak-to-promoter assignment was performed using a 10kb window with respect to the transcription starting site (TSS) of every gene on the hg19 human genome. The algorithm *msVIPER* (multi-sample Virtual Inference of Protein activity by Enriched Regulon analysis) was then applied, leveraging the full set of eight replicates per group, with 1000 signature permutations and default parameters.

qPLEX-RIME

Samples were prepared as previously described for RIME[34], protocol was modified to include TMT isobaric labels for quantification (manuscript under review). qPLEX-RIME data and analysis pipeline is available as part of the supplementary Rmarkdown.

TF Binding Overlap

Publically available data was downloaded as described in the source publication [38–40,53] and overlap was calculated with bedtools (v2.25.0). Presented data was normalised as a percentage of GRHL2 sites.

Code availability

Code for data analysis is provided as an Rmarkdown document and supporting data is avalible from https://github.com/andrewholding/VULCANSupplementary. For convenience, we have provide VULCAN as a BioConductor package https://bioconductor.org/packages/release/bioc/html/vulcan.html along with a supporting data package https://bioconductor.org/packages/release/data/experiment/html/vulcandata.html.

eRNA quantification

MCF7 cells were transfected with Smart Pool siRNA (Dharmacon, L-014515-02), siControl, GRHL2 overexpression vector (Origene, RC214498) or empty control vector using Lipofectamine 3000 (Thermo Fisher Scientific) according to the manufacturer's protocol in 6 well format. Expression was monitored by rtPCR using TaqMan assay with GAPHD as a control transcript. Knockdown efficiency was ~75% and the GRHL2 overexpression vector led a 730-fold increase in expression over control plasmid. 1 ug of purified RNA was reverse transcribed with Superscript III reverse transcriptase (Thermo Fisher Scientific, 18080085)

using random primers (Promega, C1181) according to manufacturer instructions. eRNAs were quantified with qPCR using Power SYBR™ Green PCR Master Mix (Thermo Fisher Scientific, 4367660) and denoted as relative eRNA levels after normalizing with UBC mRNA levels.

Primer name	Sequences	Reference
eGREB1 F	ACTGCGGCATTTCTGTGAGA	This study
eGREB1 R	ACTGCAGTTTGCCTGTCACT	This study
eXBP1 F	TGTGAGCACTTGGCATCCAT	Nagarajan et al 2014
eXBP1 R	ACAGGGCCTCATTCTCCTCT	Nagarajan et al 2014
eTFF1 F	AGGGGATGTGTGTGAGAAGG	Li et al 2013
eTFF1 R	GCTTCGAGACAGTGGGAGTC	Li et al 2013
UBC F	ATTTGGGTCGCGGTTCTTG	Peña et al 2009
UBC R	TGCCTTGACATTCTCGATGGT	Peña et al 2009

Author Contributions

ANH, FMG, LAS & FM conceived and designed the experiments, ANH undertook ER experimentation, ANH & FMG undertook the analysis of ER data, ANH & FMG designed VULCAN, FMG implemented the VULCAN analysis, AEC & AD undertook GRHL2

experimentation. ANH, AEC & LAS undertook analysis of GRHL2 data. ANH, FMG, LAS and FM wrote the manuscript.

Acknowledgements

We would like to acknowledge the support of the University of Cambridge, Cancer Research UK and Hutchison Whampoa Limited.

Parts of this work were funded by CRUK core grant [grant numbers C14303/A17197, A19274] to FM; Breast Cancer Now Award [grant number 2012NovPR042] to FM.

We would like to acknowledge the contribution from the CRUK Genomics, Proteomics and Bioinformatic core facilities in supporting this work.

Materials & Correspondence

All materials requests & correspondence should be addressed to Dr Andrew N. Holding.

Competing interests.

The authors have no competing interests to declare.

Abbreviations

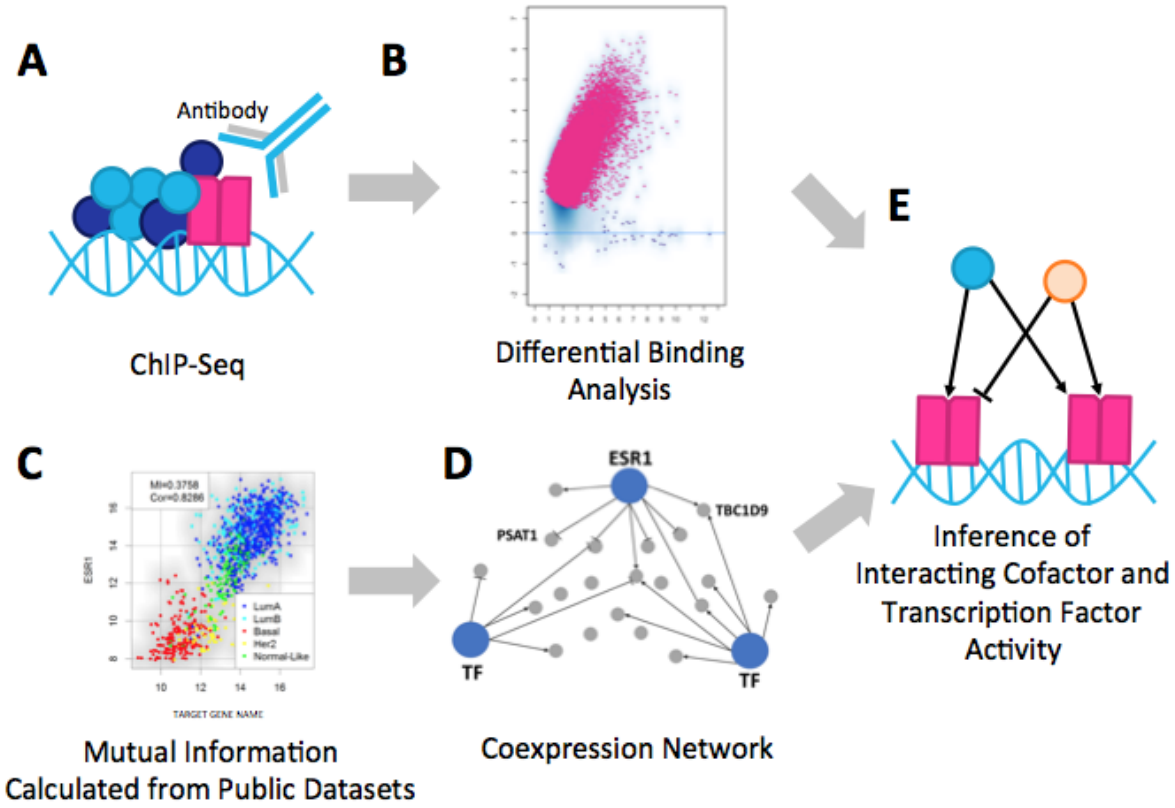
AR	Androgen Receptor
ARACNe-AP	AccuRate Algorithm for reConstruction of Network through Adaptive Partitioning
CCLE	Cancer Cell Line Encyclopedia
ChIP	Chromatin ImmunoPrecipitation
ER	Estrogen Receptor-Alpha
ERE	Estrogen-Response Elements
GSEA	Gene Set Enrichment Analysis
GRHL2	Grainyhead Like Transcription Factor 2
METABRIC	MolEcular TAxonomy of BReast cancer International Consortium
PBS	Phosphate Buffered Saline
PI	Protease Inhibitors
PR	Progesterone Receptor
TCGA	The Cancer Genome Atlas
TF	Transcription Factor
VIPER	Virtual Inference of Protein activity by Enriched Regulon analysis
VULCAN	VirtUaL ChIP-Seq Analysis through Networks

Declarations

Parts of this work were funded by CRUK core grant C14303/A17197 and A19274 (to FM), and by Breast Cancer Now Award (2012NovPR042). None of the authors have any competing interests to declare. Data supporting this manuscript is available from the GEO accession GSE109820.

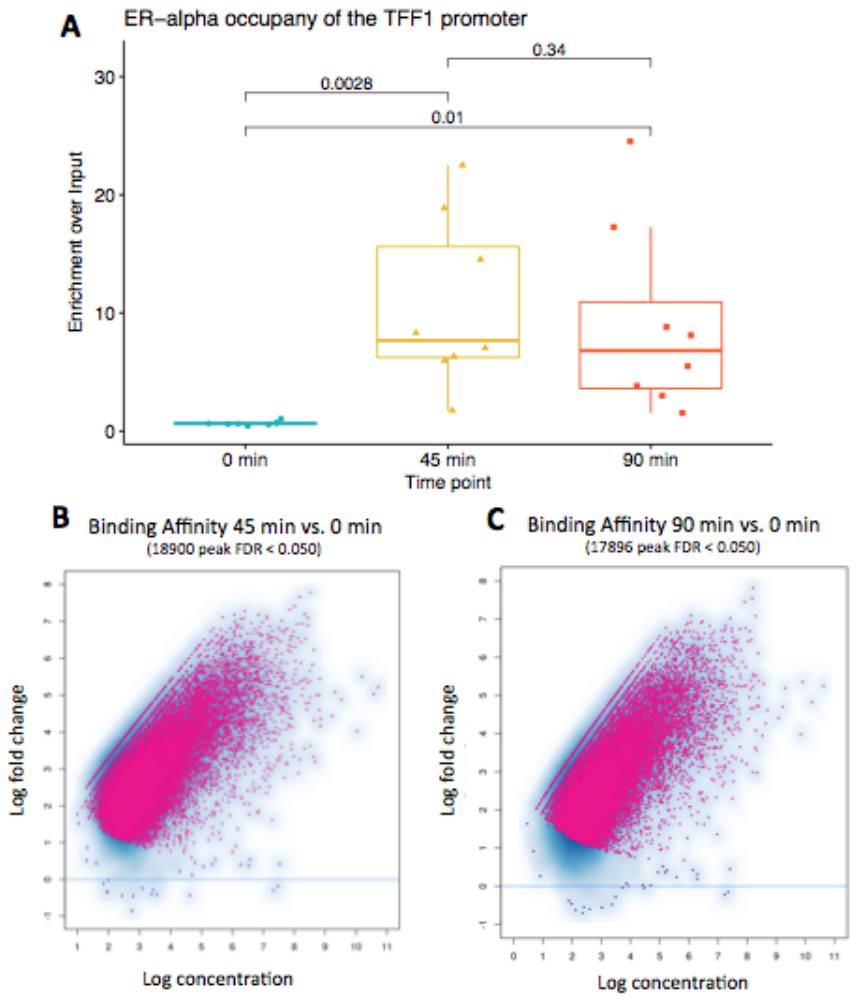
Figure Legends

Figure 1: An overview of VULCAN



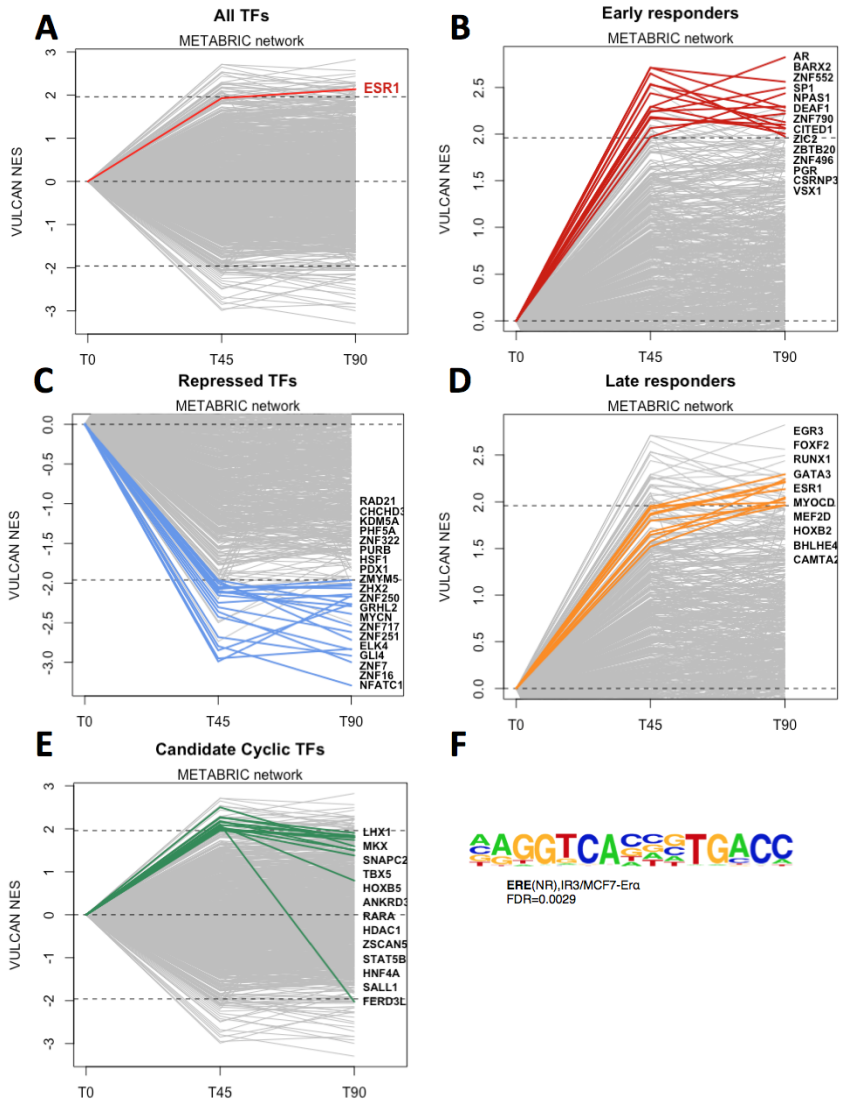
A: ChIP-Seq analysis from multiple conditions is undertaken to generate cistrome data at multiple timepoints (or conditions). B: Binding events are then compared using differential binding analysis to establish log-fold change values for individual binding events between each timepoint. C: ARACNe-AP infers all pairwise TF-target coexpression. In the example, the TCGA breast dataset is shown to infer A putative target that is correlated with ESR1. D: Minimalistic representation of the ARACNe-AP network, highlighting negative and positive regulation of targets by transcription factors. E: All the targets of a specific TF are divided in positive and negative, and tested on a differential binding signature through the msVIPER algorithm [4]

Figure 2: Dynamic behaviour during early activation of ER



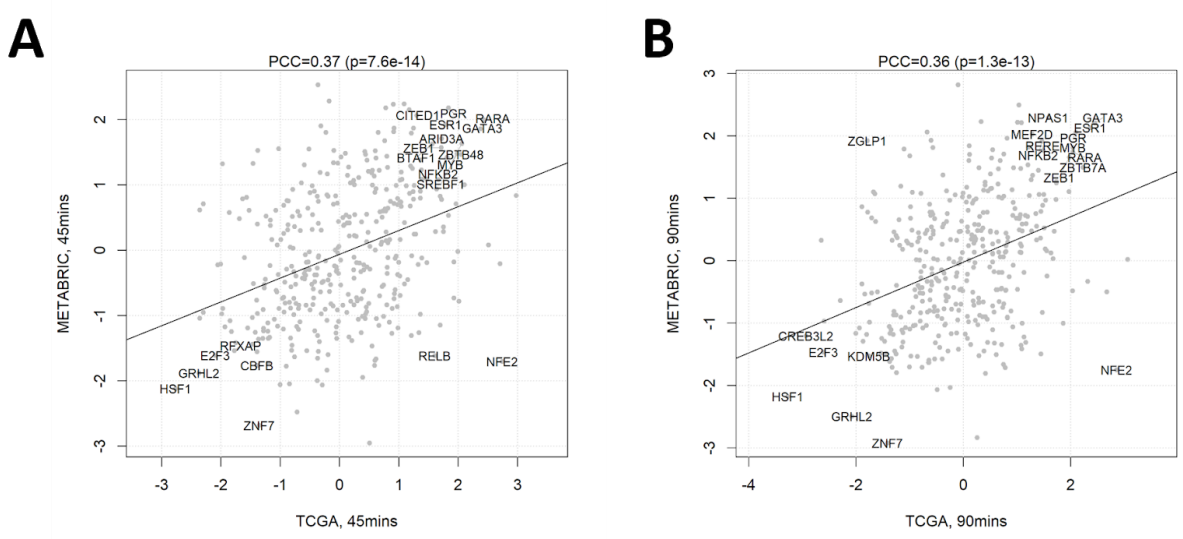
ChIP-qPCR of the TFF1 gene (A) at 3 time points shows increased binding of ER at 45 minutes after MCF7 cells are stimulated by estra-2-diol. The previously reported maximum is followed by a decrease in the TFF1 promoter occupancy at 90 minutes. P-values are generated by one-tailed t-test. The maximal point at 90 minutes was identified as an outlier (> median + 2 x IQR); however removal did not alter the significance of results. (B) Differential binding analysis of ChIP-Seq data at three time points to monitor the activation of ERa. The ER a strong increase in binding at 45 minutes vs. 0 minutes (C) and the majority of sites still display binding at 90 minutes.

Figure 3: ER occupancy after estradiol treatment in terms of TF network activity



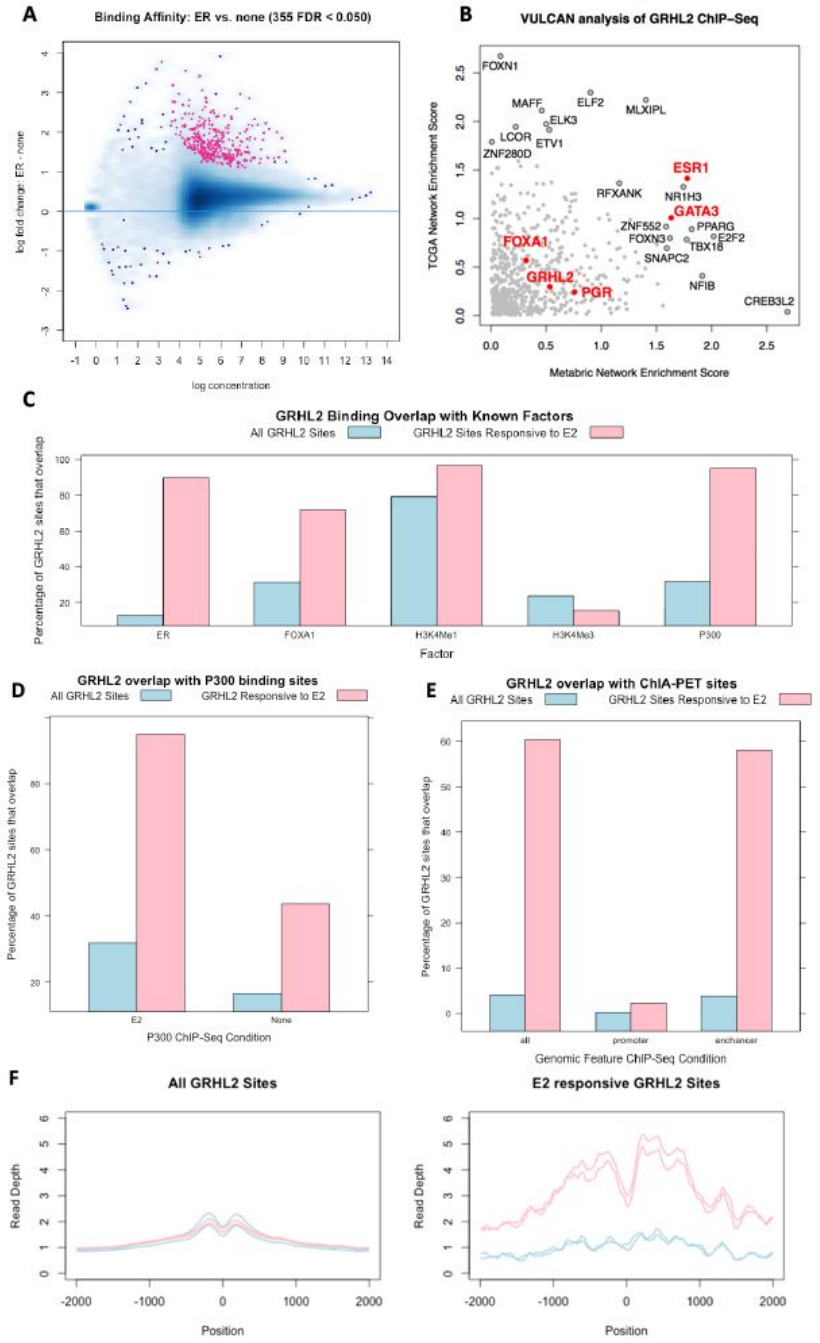
A: Global TF network behavior as predicted by VULCAN in our ChIP-Seq dataset, highlighting the ESR1 TF at time 0 and 45/90minutes after estradiol treatment.
B: Global TF activity after Estradiol Treatment in MCF7 cells, inferred using the METABRIC network, highlighting TFs significantly upregulated at 45 minutes and 90 minutes
C: Global TF activity after Estradiol Treatment in MCF7 cells, inferred using the METABRIC network, highlighting TFs significantly downregulated at 45 minutes and 90 minutes
D: Global TF activity after Estradiol Treatment in MCF7 cells, inferred using the METABRIC network, highlighting TFs significantly upregulated at 45 minutes but not at 90 minutes
E: Global TF activity after Estradiol Treatment in MCF7 cells, inferred using the METABRIC network, highlighting TFs significantly upregulated at 90 minutes but not at 45 minutes
F: Most enriched motif in peaks upregulated at both 45 and 90 minutes after estradiol treatment, as predicted by HOMER.

Figure 4: Global TF activity after estradiol treatment using different network models.



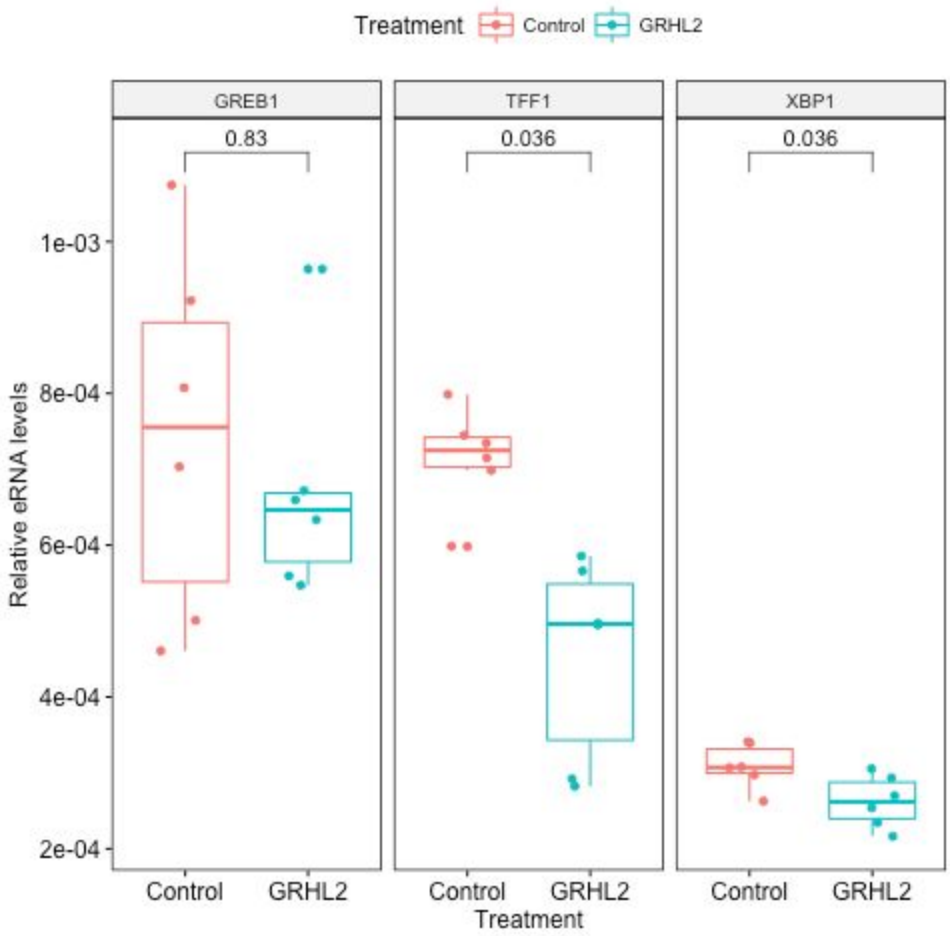
XY Scatter showing the TF activity as calculated by VULCAN for our differential ChIP-Seq analysis of ER binding at 45 minutes (A) and at 90 minutes (B) after stimulation with 100 nM E2. Comparison of results calculated using the METABRIC (y-axis) and TCGA (x-axis) networks shows consistent results know ER interactors including PGR, RARA, GATA3 and GRHL2. GRHL2 activity is notably enriched against. The regulon of ER is also consistently enriched in both networks.

Figure 5: GRHL2 Differential ChIP-Seq between 0 and 45 minutes.



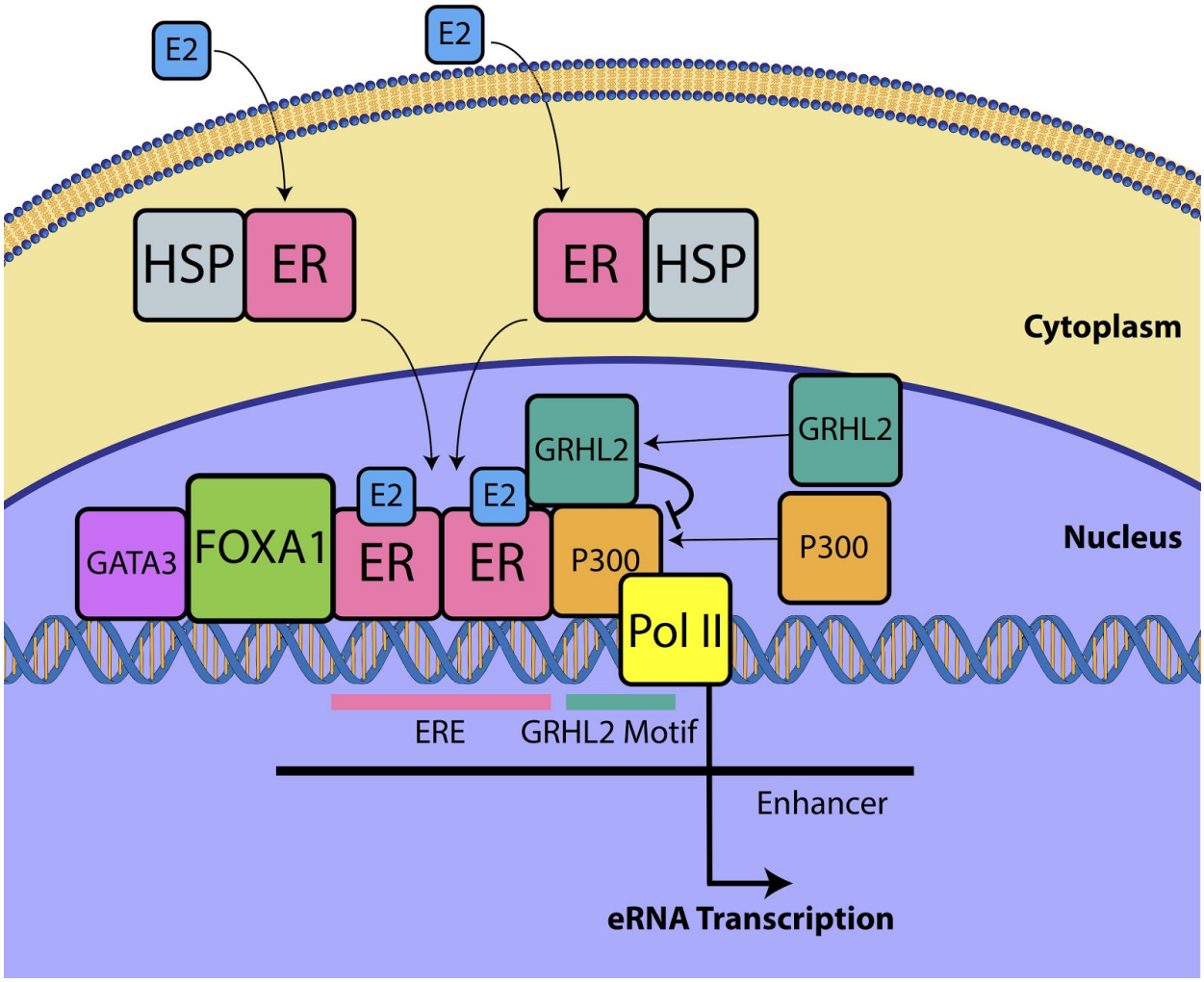
(A) Activation of the ER with estro-2-diol results in a genome wide increase in GRHL2 binding.
(B) VULCAN Analysis of the same data show a significant enrichment for ESR1 sites in both the context of the METABRIC and TGCA networks. The regulon for FOXA1 is also not enriched. Inspection of known FOXA1/GRHL2 sites (e.g. RARa promoter) shows GRHL2 already bound.
(C) Overlap of GRHL2 binding with public datasets shows that E2 responsive GRHL2 sites show considerable overlap with ER, FOXA1 and P300 sites, H3K4Me1 and H3K4Me3 show little enrichment.
(D) Analysis of P300 binding showed a greater overlap of GRHL2 ER responsive sites in the presence of E2 than in control conditions
(E) Overlap with ER ChIA-PET sites showed enrichment for GRHL2 sites at ER enhancers.
(F) Analysis of Gro-SEQ data (GSE43836) at GRHL2 sites. Blue lines are control samples, Pink are samples after stimulation with E2. In general GRHL2 sites (left) show no change in the levels of transcription on addition of E2; however, E2 responsive GRHL2 sites (right) show a robust increase in transcription on the activation of the ER.

Figure 6: Effect of overexpression of GRHL2 on eRNA at E2 responsive binding sites.



Overexpression of GRHL2 in MCF7 resulted in a reduction of eRNA transcribed from the GREB1, TFF1 and XBP1 enhancers. The effect was significant at TFF1 and XBP1 enhancers ($p < 0.05$, Wilcoxon paired-test).

Figure 7: Overview of the role of GRHL2 in ER activation



On activation of the ER by the ligand E2 the protein is released from a complex containing HSPs and translocates to the nucleus. The holo-ER dimer forms a core complex at Estrogen Response Elements (ERE) with FOXA1 (pioneer factor) and GATA3. ER further recruits P300 and GRHL2. GRHL2 has an inhibitory effect on P300 (a transcriptional activator interacting with TFIID, TFIIB, and RNAPII) thereby reducing the level of eRNA transcription at enhancer sites. Overexpression of GRHL2 further suppresses transcription, while knockdown of GRHL2 reverses the process.

References:

1. Basso K, Margolin A, Stolovitzky G, Klein U, Dalla-Favera R, Califano A. Reverse engineering of regulatory networks in human B cells. Nat Genet. Nature Publishing Group; 2005;37:382–90.
2. Giorgi FM, Del Fabbro C, Licausi F. Comparative study of RNA-seq- and Microarray-derived coexpression networks in Arabidopsis thaliana. Bioinformatics. 2013;29.
3. Castro MAA, de Santiago I, Campbell TM, Vaughn C, Hickey TE, Ross E, et al. Regulators of genetic risk of breast cancer identified by integrative network analysis. Nat. Genet. 2016;48:12–21.
4. Alvarez MJ, Shen Y, Giorgi FM, Lachmann A, Ding BB, Ye BH, et al. Functional characterization of somatic mutations in cancer using network-based inference of protein activity. Nat. Genet. 2016;48:838–47.
5. McLean CY, Bristor D, Hiller M, Clarke SL, Schaar BT, Lowe CB, et al. GREAT improves functional interpretation of cis-regulatory regions. Nat. Biotechnol. 2010;28:495–501.
6. Welch RP, Lee C, Imbriano PM, Patil S, Weymouth TE, Smith RA, et al. ChIP-Enrich: gene set enrichment testing for ChIP-seq data. Nucleic acids Res. 2014;42:e105.
7. Balwierz PJ, Pachkov M, Arnold P, Gruber AJ, Zavolan M, van Nimwegen E. ISMARA: automated modeling of genomic signals as a democracy of regulatory motifs. Genome Res. 2014;24:869–84.
8. Carroll JS. Mechanisms of oestrogen receptor (ER) gene regulation in breast cancer. Eur. J. Endocrinol. 2016;175:R41–9.
9. Lin C-Y, Vega VB, Thomsen JS, Zhang T, Kong SL, Xie M, et al. Whole-genome cartography of estrogen receptor alpha binding sites. PLoS Genet. 2007;3:e87.
10. Driscoll MD, Sathya G, Muyan M, Klinge CM, Hilf R, Bambara RA. Sequence requirements for estrogen receptor binding to estrogen response elements. J. Biol. Chem. 1998;273:29321–30.
11. Carroll JS, Meyer C a, Song J, Li W, Geistlinger TR, Eeckhoute J, et al. Genome-wide analysis of estrogen receptor binding sites. Nat. Genet. 2006;38:1289–97.
12. Carroll JS, Liu XS, Brodsky AS, Li W, Meyer CA, Szary AJ, et al. Chromosome-Wide Mapping of Estrogen Receptor Binding Reveals Long-Range Regulation Requiring the Forkhead Protein FoxA1. Cell. 2005;122.
13. Frasor J, Danes JM, Komm B, Chang KCN, Lyttle CR, Katzenellenbogen BS. Profiling of estrogen up- and down-regulated gene expression in human breast cancer cells: insights into gene networks and pathways underlying estrogenic control of proliferation and cell phenotype. Endocrinology. 2003;144:4562–74.
14. Zaret KS, Carroll JS. Pioneer transcription factors: establishing competence for gene expression. Genes & Dev. 2011;25:2227–41.
15. Horwitz KB, Jackson TA, Bain DL, Richer JK, Takimoto GS, Tung L. Nuclear receptor coactivators and corepressors. Mol. Endocrinol. 1996;10.
16. Glass CK, Rose DW, Rosenfeld MG. Nuclear receptor coactivators. Curr. Opin. cell Biol. 1997;9:222–32.
17. Green KA, Carroll JS. Oestrogen-receptor-mediated transcription and the influence of co-factors and chromatin state. Nat. Rev. Cancer. 2007;7:713–22.
18. Klinge CM. Estrogen receptor interaction with estrogen response elements. Nucleic acids Res. 2001;29:2905–19.

19. Shang Y, Hu X, DiRenzo J, Lazar MA, Brown M. Cofactor dynamics and sufficiency in estrogen receptor-regulated transcription. Cell. 2000;103:843–52.
20. Reid G, Hubner MR, Metivier R, Brand H, Denger S, Manu D, et al. Cyclic, proteasome-mediated turnover of unliganded and liganded ER alpha on responsive promoters is an integral feature of estrogen signaling. Mol. Cell. 2003;11:695–707.
21. Métivier R, Penot G, Hübner MR, Reid G, Brand H, Kos M, et al. Estrogen receptor-alpha directs ordered, cyclical, and combinatorial recruitment of cofactors on a natural target promoter. Cell. 2003;115:751–63.
22. Lachmann A, Giorgi FM, Lopez G, Califano A. ARACNe-AP: gene network reverse engineering through adaptive partitioning inference of mutual information. Bioinforma. 2016;32:2233–5.
23. Curtis C, Shah SP, Chin S-F, Turashvili G, Rueda OM, Dunning MJ, et al. The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. Nature. 2012;486:346–52.
24. Cancer Genome Atlas Network. Comprehensive molecular portraits of human breast tumours. Nature. 2012;490:61–70.
25. Lefebvre C, Rajbhandari P, Alvarez MJ, Bandaru P, Lim WK, Sato M, et al. A human B-cell interactome identifies MYB and FOXM1 as master regulators of proliferation in germinal centers. Mol. Syst. Biol. 2010;6:377.
26. Lachmann A, Giorgi FM, Lopez G, Califano A. ARACNe-AP: gene network reverse engineering through adaptive partitioning inference of mutual information. Bioinformatics. 2016;32.
27. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc. Natl. Acad. Sci. United States Am. 2005;102:15545–50.
28. Orlando DA, Chen MW, Brown VE, Solanki S, Choi YJ, Olson ER, et al. Quantitative ChIP-Seq normalization reveals global modulation of the epigenome. Cell reports. 2014;9:1163–70.
29. Guertin MJ, Markowetz F, Holding AN. Novel Quantitative ChIP-seq Methods Measure Absolute Fold-Change in ER Binding Upon Fulvestrant Treatment. bioRxiv. 2017;
30. Ross-Innes CS, Stark R, Teschendorff AE, Holmes KA, Ali HR, Dunning MJ, et al. Differential oestrogen receptor binding is associated with clinical outcome in breast cancer. Nature. 2012;481:389–93.
31. Stein RA, Chang C-Y, Kazmin DA, Way J, Schroeder T, Wergin M, et al. Estrogen-related receptor alpha is critical for the growth of estrogen receptor-negative breast cancer. Cancer Res. 2008;68:8805–12.
32. Bhat-Nakshatri P, Wang G, Appaiah H, Luktuke N, Carroll JS, Geistlinger TR, et al. AKT alters genome-wide estrogen receptor alpha binding and impacts estrogen signaling in breast cancer. Mol. Cell. Biol. 2008;28:7487–503.
33. Gozgit JM, Pentecost BT, Marconi SA, Ricketts-Loriaux RSJ, Otis CN, Arcaro KF. PLD1 is overexpressed in an ER-negative MCF-7 cell line variant and a subset of phospho-Akt-negative breast carcinomas. Br. J. Cancer. 2007;97:809–17.
34. Mohammed H, D'Santos C, Serandour AA, Ali HR, Brown GD, Atkins A, et al. Endogenous purification reveals GREB1 as a key estrogen receptor regulatory factor. Cell reports. 2013;3:342–9.
35. Thompson A, Schäfer J, Kuhn K, Kienle S, Schwarz J, Schmidt G, et al. Tandem Mass Tags: A Novel Quantification Strategy for Comparative Analysis of Complex Protein Mixtures by

MS/MS. Anal. Chem. 2003;75.
36. Xiang X, Deng Z, Zhuang X, Ju S, Mu J, Jiang H, et al. Grhl2 determines the epithelial phenotype of breast cancers and promotes tumor progression. PloS one. 2012;7:e50781.
37. Cieply B, Riley P, Pifer PM, Widmeyer J, Addison JB, Ivanov AV, et al. Suppression of the epithelial-mesenchymal transition by Grainyhead-like-2. Cancer Res. 2012;72:2440–53.
38. Jozwik KM, Chernukhin I, Serandour AA, Nagarajan S, Carroll JS. FOXA1 Directs H3K4 Monomethylation at Enhancers via Recruitment of the Methyltransferase MLL3. Cell reports. 2016;17:2715–23.
39. Zwart W, Theodorou V, Kok M, Canisius S, Linn S, Carroll JS. Oestrogen receptor-co-factor-chromatin specificity in the transcriptional regulation of breast cancer. EMBO J. 2011;30:4764–76.
40. Hah N, Murakami S, Nagari A, Danko CG, Kraus WL. Enhancer transcripts mark active estrogen receptor binding sites. Genome Res. 2013;23:1210–23.
41. Pifer PM, Farris JC, Thomas AL, Stoilov P, Denvir J, Smith DM, et al. Grainyhead-like 2 inhibits the coactivator p300, suppressing tubulogenesis and the epithelial-mesenchymal transition. Mol. Biol. cell. 2016;27:2479–92.
42. Paltoglou S, Das R, Townley SL, Hickey TE, Tarulli GA, Coutinho I, et al. Novel Androgen Receptor Coregulator GRHL2 Exerts Both Oncogenic and Antimetastatic Functions in Prostate Cancer. Cancer Res. 2017;77:3417–30.
43. Cheng J-H, Pan DZ-C, Tsai ZT-Y, Tsai H-K. Genome-wide analysis of enhancer RNA in gene regulation across 12 mouse tissues. Sci. reports. 2015;5:12648.
44. Rahman S, Zorca CE, Traboulsi T, Noutahi E, Krause MR, Mader S, et al. Single-cell profiling reveals that eRNA accumulation at enhancer-promoter loops is not required to sustain transcription. Nucleic acids Res. 2017;45:3017–30.
45. Kim T-K, Hemberg M, Gray JM, Costa AM, Bear DM, Wu J, et al. Widespread transcription at neuronal activity-regulated enhancers. Nature. 2010;465:182–7.
46. Holmes KA, Brown GD, Carroll JS. Chromatin Immunoprecipitation-Sequencing (ChIP-seq) for Mapping of Estrogen Receptor-Chromatin Interactions in Breast Cancer. Methods Mol. Biol. 2016;1366:79–98.
47. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics. 2009;25.
48. Zhang Y, Liu T, Meyer CA, Eeckhoute J, Johnson DS, Bernstein BE, et al. Model-based analysis of ChIP-Seq (MACS). Genome Biol. 2008;9:137.
49. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles: a knowledge-based approach for interpreting genome-wide expression profiles. Proc. Natl. Acad. Sci. United States Am. National Acad Sciences; 2005;102:15545–50.
50. Heinz S, Benner C, Spann N, Bertolino E, Lin YC, Laslo P, et al. Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. Mol. cell. 2010;38:576–89.
51. Crooks GE, Hon G, Chandonia J-M, Brenner SE. WebLogo: a sequence logo generator. Genome Res. 2004;14:1188–90.
52. Anders S, Huber W. Differential expression analysis for sequence count data. Genome Biol. BioMed Central Ltd; 2010;11:106.
53. Hurtado A, Holmes KA, Ross-Innes CS, Schmidt D, Carroll JS. FOXA1 is a key determinant of estrogen receptor function and endocrine response. Nat Genet. Nature Publishing Group; 2010;43:27–33.