A Appendices

A.1 AuROC and AuPRC for identifying TAL motif

Similar to Fig. 4, the auROC and auPRC curves for identifying the TAL motif using importance scores are shown in fig A.1.
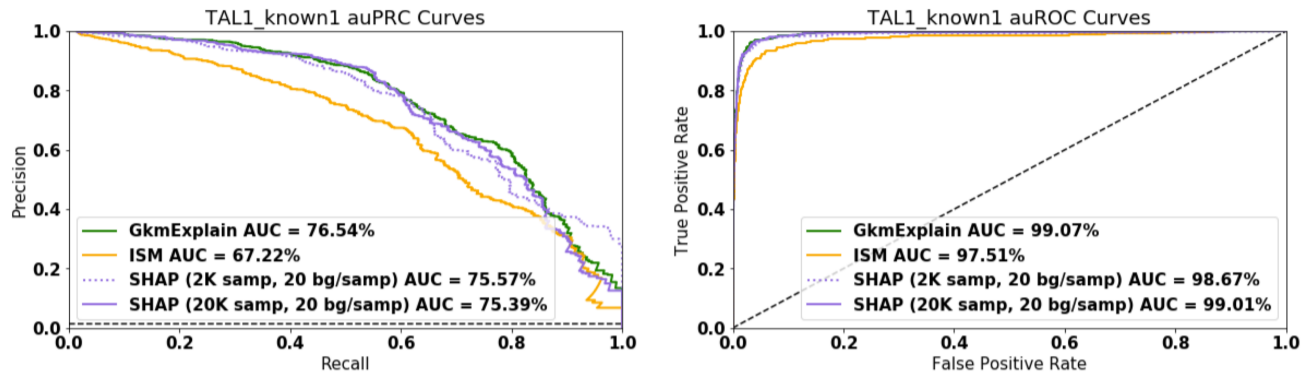

Fig. A.1: **GkmExplain outperforms ISM at identifying TAL1 motifs**. Analogous to **Fig. 4**, but using 16bp windows (the length of the TAL1_known1 motif) rather than 10bp windows.

A.2 Connection to Integrated Gradients

This section elucidates the theoretical connections between the formulas in **Sec. 5.1** and the method of Integrated Gradients [16].

For any output function $F(X)$, where X is a vector in $\mathbb{R}^n$, and a path function $\pi : [0,1] \to \mathbb{R}^n$, the *Path Integrated Gradients* [16] importance on the ith feature X_i is defined as:

$$\text{PathIntegratedGradients}_i^{\pi} ::= \int_{\beta=0}^{\beta=1} \frac{\partial F(\pi(\beta))}{\partial \pi_i(\beta)} \frac{d\pi_i(\beta)}{d\beta} d\beta \tag{18}$$

Applying this definition to the objective function for SVMs in **Eqn. 1** we get:

$$\text{SVM\_PIG}_i^{\pi} ::= \int_{\beta=0}^{\beta=1} \sum_{j=1}^{m} \alpha_j y^j \frac{\partial K(Z^j, \pi(\beta))}{\partial \pi_i(\beta)} \frac{d\pi_i(\beta)}{d\beta} d\beta \tag{19}$$

The primary hurdle when applying Integrated Gradients to support vector machine string kernels is that the formula for Integrated Gradients requires the kernel function K to be differentiable. Because DNA sequence is discrete, most string kernels are not differentiable. To circumvent this issue, we introduce vector of variables η^{S_x} where each element $\eta_i^{S_x}$ can be interpreted as the "intensity" of the base at position i of sequence S_x. Let S_x^i denote the identity of base i in sequence S_x, let l be the length of the l-mers, and let $1\{S_x^i = S_x^j\}$ be an indicator function that takes the value of 1 if $S_x^i = S_x^j$ and 0 otherwise. We rewrite the wgkm kernel between an input sequence S_x and a support vector sequence S_z as a function of η^{S_x} as follows:

$$K_{\text{wgkm}}^{(S_x, S_z)}(\eta^{S_x}) = \sum_i \sum_j \left(\frac{w_i w_j h(f_m(u_i^{S_x}, u_j^{S_z}))}{\left\| f_{\text{wgkm}}^{S_x} \right\| \left\| f_{\text{wgkm}}^{S_z} \right\|} \sum_{k=0}^{l-1} \frac{\eta_{i+k}^{S_x} 1\{S_x^{i+k} = S_z^{j+k}\}}{l - f_m(u_i^{S_x}, u_j^{S_z})} \right) \tag{20}$$

When $\eta_i^{S_x} = 1$ for all i, $K_{\text{wgkm}}^{(S_x, S_z)}(\eta^{S_x})$ is equal to $K_{\text{wgkm}}(S_x, S_z)$.

Recall that $w_i w_j h(f_m(u_i^{S_x}, u_j^{S_z}))$ is the contribution of the l-mer pair $(u_i^{S_x}, u_j^{S_z})$ to $\langle f_{\text{wgkm}}^{S_x}, f_{\text{wgkm}}^{S_z} \rangle$. In the expression above, $w_i w_j h(f_m(u_i^{S_x}, u_j^{S_z}))$ is scaled according to the total intensity of all the $l - f_m(u_i^{S_x}, u_j^{S_z})$ matching bases between $u_i^{S_x}$ and $u_j^{S_z}$.

Let $\text{len}(S_x)$ be a function that returns the length of the sequence S_x. Because **Eqn. 20** is differentiable, we can compute the partial derivative $\frac{\partial K_{\text{wgkm}}^{(S_x, S_z)}(\eta^{S_x})}{\partial \eta_i^{S_x}}$ as:

$$\frac{\partial K_{\text{wgkm}}^{(S_x, S_z)}(\eta^{S_x})}{\partial \eta_i^{S_x}} = \sum_{j=\max(i-(l-1),0)}^{\min(i,\text{len}(S_x)-l)} \sum_k \left(\frac{w_k w_j h(f_m(u_j^{S_x}, u_k^{S_z}))}{\|f^{S_x}\| \|f^{S_z}\|} \frac{1\{S_x^i = S_z^{k+(i-j)}\}}{l - f_m(u_j^{S_x}, u_k^{S_z})} \right) \tag{21}$$

Intuitively, the expression looks at each l-mer $u_j^{S_x}$ in S_x that overlaps position i, compares it to every l-mer $u_k^{S_z}$ in S_z, and increments the gradient proportionally to $w_i w_j h(f_m(u_j^{S_x}, u_k^{S_z}))$ only if the base at position $(i-j)$ in $u_j^{S_x}$ is a match to the corresponding base in $u_k^{S_z}$. Note that the number of l-mers in sequence S_x is $(\text{len}(S_x) - (l-1))$

Rewriting the formula for Path Integrated Gradients in **Eqn. 19** using the differentiable reparameterization in **Eqn. 20**, we get:

$$\text{wgkm\_SVM\_PIG}_i^{(\pi, S_x)} ::= \int_{\beta=0}^{\beta=1} \sum_{j=1}^{m} \alpha_j y^j \frac{\partial K_{\text{wgkm}}^{(S_z^j, S_x)}(\pi(\beta))}{\partial \pi_i(\beta)} \frac{d\pi_i(\beta)}{d\beta} d\beta \tag{22}$$

If we now consider the path $\pi_i(\beta) = \beta$ (i.e. if we linearly scale the intensities $\eta_i^{S_x}$ uniformly between 0 and 1), we would have $\frac{d\pi_i(\beta)}{d\beta} = 1$. This choice of path is the standard "Integrated Gradients" path. Because the partial derivative in **Eqn. 21** is constant with respect to $\eta_i^{S_x}$, simplifying the expression above results in the formula for the importance of individual bases in a wkgm SVM that was presented earlier in **Eqn. 11**, reproduced below for convenience:

$$\text{wgkm\_SVM\_IG}_i^{S_x} = \phi_{i,S_x}^{\text{wgkmsvm}} = \sum_{j=1}^{m} \alpha_j y^j \phi_{i,S_x,S_{z^j}}^{\text{wgkm}} \tag{23}$$

Where S_{z^j}, α_j and y^j are, respectively, the sequence, weight and label of support vector j.

For the case of the wgkmrbf kernel (**Eqn. 7**), we once again reparameterize the kernel in terms of $\eta_i^{S_x}$ as follows:

$$K_{\text{wgkmrbf}}^{(S_x, S_z)}(\eta^{S_x}) = \exp\left(\gamma(K_{\text{wgkm}}^{(S_x, S_z)}(\eta^{S_x}) - 1) \right) \tag{24}$$

By the chain rule, we get:

$$\frac{\partial K_{\text{wgkmrbf}}^{(S_x, S_z)}(\eta^{S_x})}{\partial \eta_i^{S_x}} = \gamma \frac{\partial K_{\text{wgkm}}^{(S_x, S_z)}(\eta^{S_x})}{\partial \eta_i^{S_x}} \exp\left(\gamma(K_{\text{wgkm}}^{(S_x, S_z)}(\eta^{S_x}) - 1) \right) \tag{25}$$

As before, we consider the path $\pi_i(\beta) = \beta$ (i.e. we linearly scale the intensities $\eta_i^{S_x}$ uniformly between 0 and 1). From **Eqn. 20**, we note that $K_{\text{wgkm}}^{(S_x, S_z)}(\eta^{S_x})$ is linear in the values of $\eta_i^{S_x}$. Thus, we can write:

$$\text{wgkmrbf\_SVM\_IG}_i^{S_x} = \int_{\beta=0}^{\beta=1} \sum_{j=1}^{m} \alpha_j y^j \gamma \frac{\partial K_{\text{wgkm}}^{(S_x, S_z)}(\eta^{S_x})}{\partial \pi_i(\beta)} \exp\big(\gamma \beta K_{\text{wgkm}}(S_x, S_z) - \gamma\big) d\beta$$

Recall that $\frac{\partial K_{\text{wgkm}}^{(S_x, S_z)}(\eta^{S_x})}{\partial \pi_i(\beta)}$ is constant with respect to $\pi_i(\beta)$. After performing the integral with respect to β, we obtain **Eqn. 13**, reproduced below for convenience:

$$\text{wgkmrbf\_SVM\_IG}_i^{S_x} = \phi_{i,S_x}^{\text{wgkmrbfsvm}} = \sum_{j=1}^{m} \alpha_j y^j \phi_{i,S_x}^{\text{wgkmrbf}} \tag{26}$$

A.3 Normalization of scores for TF-MoDISco

Empirically, we found that the following normalization of the importance scores produces improved results with TF-MoDISco. Let $f_h(S_x, i, B^*)$ be a function that returns the hypothetical importance of base B^* at position i in sequence S_x, and let S_x^i be the original base at position i in sequence S_x. To normalize the hypothetical importance scores at position i, we divide by the sum of all hypothetical scores at position i with the same sign as $f_h(S_x, i, S_x^i)$. The rationale is that if a different base at position i could produce a score of higher magnitude than S_x^i, then S_x^i is relatively less important. Let $1\{x > 0\}$ be an indicator function that returns 1 if x is positive and 0 otherwise. Formally, the normalized hypothetical importance for base B^* at position i is defined as:

$$\bar{f}_h(S_x, i, B^*) ::= \frac{f_h(S_x, i, B^*)f_h(S_x, i, S_x^i)}{\sum_{B'} f_h(S_x, i, B')1\{(f_h(S_x, i, B')f_h(S_x, i, S_x^i)) > 0\}} \tag{27}$$

Similarly, we define the normalized importance score as the value of the normalized hypothetical importance score for the base that is actually present in the original sequence:

$$\bar{f}_i(S_x, i) ::= \bar{f}_h(S_x, i, S_x^i) \tag{28}$$

We find that the normalized importance scores appear to be less noisy relative to the unnormalized importance scores, as illustrated in **Fig. A.2**.
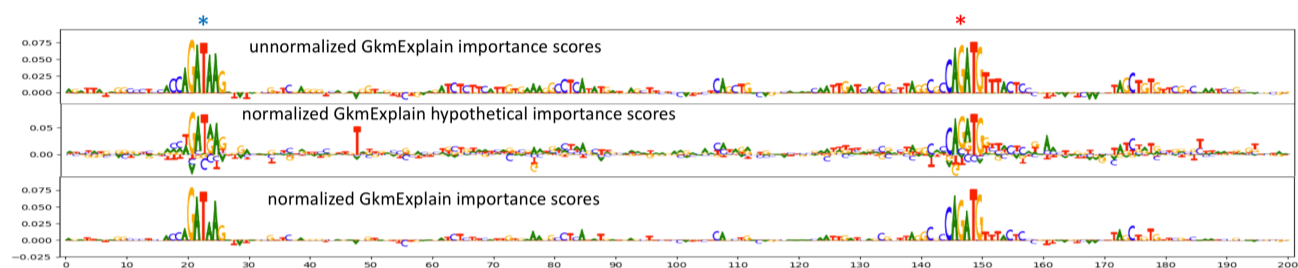

Fig. A.2: **Normalization of GkmExplain importance scores for TF-MoDISco**. A gkmrbf SVM was trained as described in **Sec. 6.1**. Shown are unnormalized GkmExplain importance scores (top row), normalized hypothetical importance scores (**Eqn. 27**), and normalized importance scores (**Eqn. 28**) on a single sequence. The blue star indicates the location of an embedded GATA1 motif (GATAAG), and the red star indicates the location of an embedded TAL1 motif (CAGATG). The normalized importance scores appear less noisy than the unnormalized importance scores.

A.4 AuROC on dsQTL data

Performance in terms of auROC on the dsQTL dataset is shown in **Fig. A.3**. Although GkmExplain gives a weaker auROC than ISM in 4 out of 5 cases, the difference is not statistically significant by a Wilcoxon test (possibly owing to the small number of samples). We note that when negatives greatly outnumber positives, auPRC is generally considered a more useful performance metric than auROC. We also note that GkmExplain consistently outperforms deltaSVM.
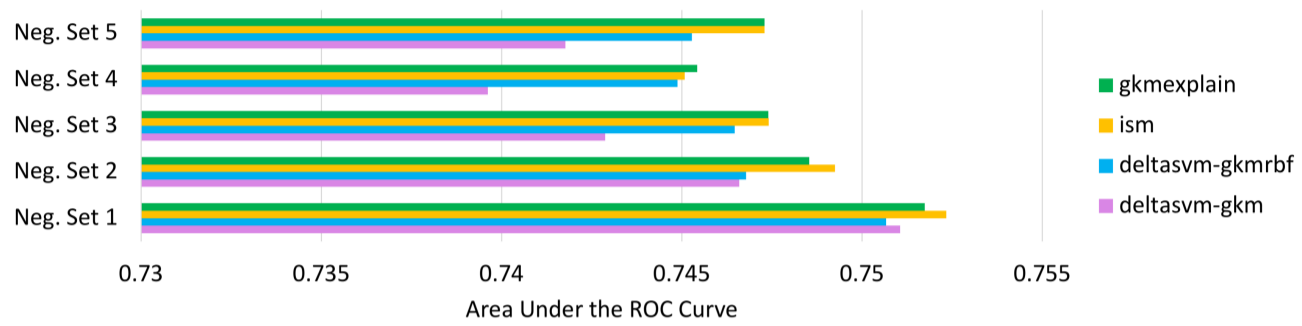

Fig. A.3: **AuROC for dsQTL identification**. Analogous to **Fig. 8**, but showing auROC instead of auPRC. The GkmExplain and ISM methods consistently outperform deltaSVM.

A.5 Integrated Gradients for interpreting other non-linear SVMs

To explore the general applicability of Integrated Gradients (IG) to SVM interpretation beyond genomics, we used the MADELON dataset [1]. This dataset has a training set of 2000 examples and a validation set of 600 examples, all with 500 features, of which 20 features are informative and the remaining are "distractor" features with no predictive power. Because the Gaussian SVM had a strong tendency to overfit to the full dataset, we created 10 reduced feature sets, each with the same set of examples in the training and validation sets, but with 20 likely informative features (identified based on the challenge submission results) and a different set of 20 distractor features in each set. Each of these 10 sets served as a distinct data point on which to evaluate the accuracy of the importance scoring algorithm in question (Integrated Gradients or SHAP). Each feature in each set was normalized by subtracting the average feature value and dividing by the standard deviation. We trained Gaussian kernel SVM on each of the datasets and achieved approximately 0.8 accuracy on the validation set. These classifiers were then provided to the importance scoring algorithms.

When applying Integrated Gradients, we used a reference that was the average feature value of the negative labeled training points and approximated the integration of the partial derivatives using 10 linearly-spaced intermediate points between each test point and the reference. With SHAP, we experimented with 10 and 50 samples around the test points using the same reference as was used for IG. We judged the quality of each importance scoring algorithm by feeding the top 10 most important features reported by the algorithm to a Gaussian kernel SVM and measuring the SVM's accuracy on the validation set. The results are shown in Figure A.4. Integrated Gradients is significantly faster than SHAP and tends to be more accurate than when SHAP is used with a small number of perturbation samples. With a larger number of perturbation samples, SHAP is more accurate, though it is still slower. A notebook demonstrating how to use Integrated Gradients on a non-genomic dataset is at:
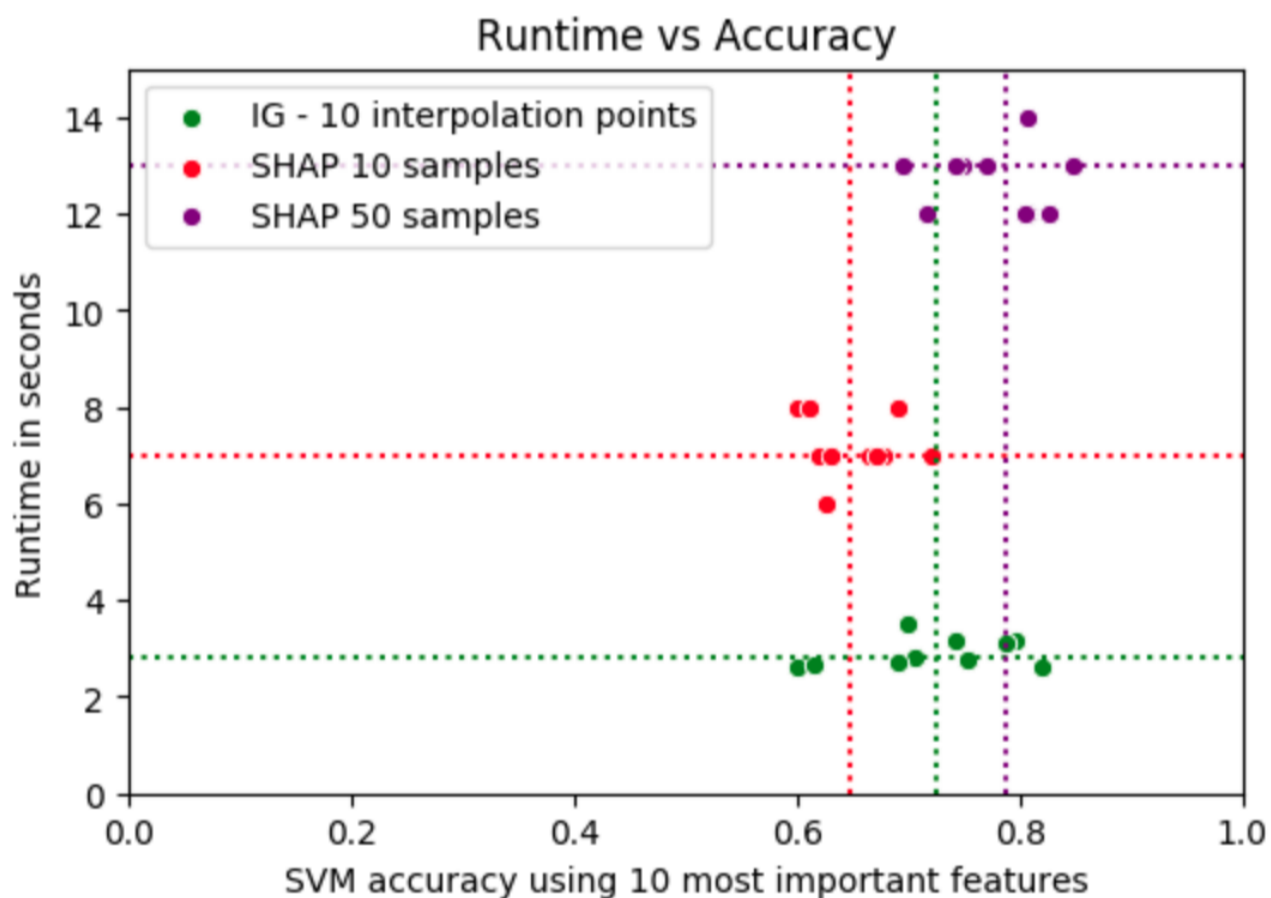
`https://colab.research.google.com/drive/1LUGMIwOHLKDdK3deg7IIZ71kb2Nu1qv2`.

Fig. A.4: **Comparison of Integrated Gradients with SHAP for Gaussian kernel SVM interpretation on the Madelon dataset**. Dashed lines indicate means. IG appears to be pareto optimal in the sense that it can provide more accurate results faster than compared to SHAP used with a small number of perturbation samples.

References

[1]Uci machine learning repository: Madelon data set. `https://archive.ics.uci.edu/ml/datasets/Madelon`. (Accessed on 08/01/2018).
[2]www.beerlab.org/deltasvm/. `http://www.beerlab.org/deltasvm/`. (Accessed on 10/26/2018).

[3] Timothy L Bailey, Mikael Boden, Fabian A Buske, Martin Frith, Charles E Grant, Luca Clementi, Jingyuan Ren, Wilfred W Li, and William S Noble. MEME SUITE: tools for motif discovery and searching. *Nucleic Acids Res.*, 37(Web Server issue):W202–8, July 2009.

[4] Yana Bromberg and Burkhard Rost. Comprehensive in silico mutagenesis highlights functionally important residues in proteins. *Bioinformatics*, 24(16):i207–12, August 2008.

[5] Jacob F Degner, Athma A Pai, Roger Pique-Regi, Jean-Baptiste Veyrieras, Daniel J Gaffney, Joseph K Pickrell, Sherryl De Leon, Katelyn Michelini, Noah Lewellen, Gregory E Crawford, Matthew Stephens, Yoav Gilad, and Jonathan K Pritchard. DNase I sensitivity QTLs are a major determinant of human expression variation. *Nature*, 482(7385):390–394, February 2012.

[6] ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature*, 489(7414):57–74, September 2012.

[7] Mahmoud Ghandi, Dongwon Lee, Morteza Mohammad-Noori, and Michael A. Beer. Enhanced regulatory sequence prediction using gapped k-mer features. *PLoS Comput. Biol.*, 10(7):e1003711, jul 2014.

[8] Sven Heinz, Christopher Benner, Nathanael Spann, Eric Bertolino, Yin C Lin, Peter Laslo, Jason X Cheng, Cornelis Murre, Harinder Singh, and Christopher K Glass. Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Mol. Cell*, 38(4):576–589, May 2010.

[9] Pouya Kheradpour and Manolis Kellis. Systematic discovery and characterization of regulatory motifs in ENCODE TF binding experiments. *Nucleic Acids Res.*, 42(5):2976–2987, March 2014.

[10] Dongwon Lee. LS-GKM: a new gkm-SVM for large-scale datasets. *Bioinformatics*, 32(14):2196–2198, July 2016.

[11] Dongwon Lee, David U Gorkin, Maggie Baker, Benjamin J Strober, Alessandro L Asoni, Andrew S McCallion, and Michael A Beer. A method to predict the impact of regulatory variants from DNA sequence. *Nat. Genet.*, 47(8):955–961, August 2015.

[12] Christina Leslie and Rui Kuang. Fast String Kernels using Inexact Matching for Protein Sequences. *J. Mach. Learn. Res.*, 5(Nov):1435–1455, 2004.

[13] Scott Lundberg and Su-In Lee. A unified approach to interpreting model predictions. *NIPS*, 2017.

[14] Avanti Shrikumar, Peyton Greenside, and Anshul Kundaje. Learning important features through propagating activation differences. *CoRR*, abs/1704.02685, 2017.

[15] Avanti Shrikumar, Katherine Tian, Anna Shcherbina, Žiga Avsec, Abhimanyu Banerjee, Mahfuza Sharmin, Surag Nair, and Anshul Kundaje. Tf-modisco v0.4.2.2-alpha: Technical note. *CoRR*, abs/1811.00416.

[16] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. *CoRR*, abs/1703.01365, 2017.

[17] Jian Zhou and Olga G Troyanskaya. Predicting effects of noncoding variants with deep learning–based sequence model. *Nat. Methods*, 12(10):931–934, August 2015.