Comparing peaks from pseudo-bulk ATAC-seq and H3K27ac ChIP-seq

Introduction and Summary of Main Results

Background and motivation. ATAC-seq signal, indicating DNA accessibility, and H3K27ac ChIP-seq signal, a sign of active enhancers and promoters, are important for determining active regulatory regions. Subsequently, we can find and study the genes controlled by these regulatory regions. A key motivation is to understand how the peaks from these 2 different techniques compare - which is more informative? Could combining both be more informative? Moreover, using these peaks can provide insights into GWAS SNPs linked to diseases like Alzheimer's.

In this project we consider the following **key research questions** and obtain these **findings**:

1. Is the difference between ChIP-ATAC-shared and ChIP-specific or pseudo-bulk ATAC-specific peaks large enough to be predictable by a machine learning classifier?
   Yes, the differences are indeed predictable, at an auROC of ~85-90%.
2. Are there properties which distinguish the overlapping (H3K27ac ChIP and ATAC) peaks and the unique peaks (either type)?
   Properties which distinguish overlapping versus unique include: bigWig signal of peak, GC content, distance from enhancer, chromatin state proportions.
3. Can we find possible AD GWAS SNPs from accessibility data?
   Only preliminary analysis done.

The following is an outline of **key methodological choices and approaches** used:

- Data Collection: gather relevant information from input narrowPeak files
- Processing each peak (ChIP or ATAC) to find whether they overlapped with the other type of peak, and find features[1].
- Balance data to obtain similar size positive (overlapping) and negative (unique) training samples for classifier: used an automated way to choose the best balancing technique[2].
- Cross validation testing to find best hyperparameters (automated) for the different ML models[3]: e.g. best *maxnodes* value for random forest.
- Feature selection, to choose (automated) which features gave the best classification.
- The best model out of the four[4] used chosen by human judgement.
- Later, a gapped k-mer SVM (gkmSVM) model (from Kundaje lab) was trained using ATAC accessible (positive set) and non-accessible DNA (negative set), to find important "implicit" sequence features; these were then used to find importance of single nucleotide polymorphisms (SNPs) in DNA, and whether it significantly changed an important motif, possibly of a transcription factor (TF).

Activities during the internship:

- *Week 1-3: reading papers*
- *Week 4-5: understanding pipelines (ENCODE ChIP-seq pipeline, ArchR for scATAC), and testing them out*
- *Week 6-12: coding for preprocessing, doing analysis on comparing peaks, making machine learning classifiers, and learning/working on gkmSVM pipeline.*
- *Week 13-14: continuing coding (part time)*

[1] bigWig signal, GC content, genomic distances (e.g. from promoters or enhancers), chromatin states
[2] Chosen from the following: oversampling, SMOTE oversampling, undersampling, and no balancing
[3] The different models tested were logistic regression, random forest, XGBoost, and support vector machine (SVM).
[4] Logistic regression, random forest, XGBoost, SVM

Is the difference between ChIP-ATAC-shared and ChIP-specific or ATAC-specific peaks large enough to be predictable by a machine learning classifier?

A machine learning classifier was trained to predict whether an H3K27ac ChIP peak was overlapping or not (was unique) with a ATAC peak, and vice versa. Various features were used to understand the important properties of the peaks such as GC content, signal, chromatin state, distances to genes and enhancers, etc. GC content is known to be higher around promoter regions, so was hypothesised to be a suitable basis for separating ATAC and H3K27ac ChIP overlapping and unique peaks; similarly for distance to genes[5] (shorter distances would be promoter or possibly intronic enhancers) and distance to enhancers[6]. Signal would be important to find the most significant peaks, which would hopefully be higher in the overlapping peaks. Moreover chromatin states[7] were used as a final feature, which could potentially give a better understanding of the types of chromatin where H3K27ac modifications versus ATAC accessible regions would be at.

Using H3K27ac ChIP to predict overlap with scATAC (chr21)			**Using scATAC to predict overlap with H3K27ac ChIP (chr21)**		
Percentages		Performance	Percentages		Performance
auROC baseline = **50%**	logistic	**86.33%**	auROC baseline = **50%**	logistic	86.19%
	random.forest	86.59%		random.forest	**89.81%**
	xgboost	84.41%		xgboost	87.55%
	svm	**86.66%**		svm	88.41%
F1 Score baseline = **35%**	logistic	**69.06%**	F1 Score baseline = **61%**	logistic	89.24%
	random.forest	68.14%		random.forest	**90.66%**
	xgboost	68.25%		xgboost	**90.51%**
	svm	**69.92%**		svm	90.05%
auPRC baseline = **73%**	logistic	**93.26%**	auPRC baseline = **21%**	logistic	56.98%
	random.forest	**93.67%**		random.forest	**66.36%**
	xgboost	91.85%		xgboost	61.22%
	svm	**93.52%**		svm	61.35%

Figure 1: performance on test data, using peaks from a sample, in this case chromosome 21 (Left: using H3K27ac ChIP to predict overlap with ATAC, Right: using ATAC to predict overlap with H3K27ac ChIP)

The results affirm that these differences are predictable using the aforementioned features; the auROC (area under receiver operating characteristic curve) increases from a 50% baseline to ~85-90%. Similar increases were found for auPRC (area under precision-recall curve) and F1 score (Figure 1 & Appendix 1). The best models for ChIP (overlapping with ATAC v.s. unique) and ATAC (overlapping with ChIP v.s. unique) data were logistic regression and random forest, respectively.

Method. A systematic and sequential method was used to improve the classifier (shown in Figure 2). First, the best technique for balancing the input data (e.g. simple oversampling) was determined, then the best hyperparameters for each model were chosen (e.g. ntree=300 for random forest), and finally the best features were selected (e.g. using all chromatin states and signal data). Performance improves from a baseline of 50% to ~85-90% as shown in Appendix 1.

[5] annotations taken from GENCODE
[6] annotations for macrophages (cell type judged to be closest to microglia) taken from EnhancerAtlas
[7] taken from the 18 chromatin state ChromHMM model for monocytes (cell type judged to be closest to microglia)

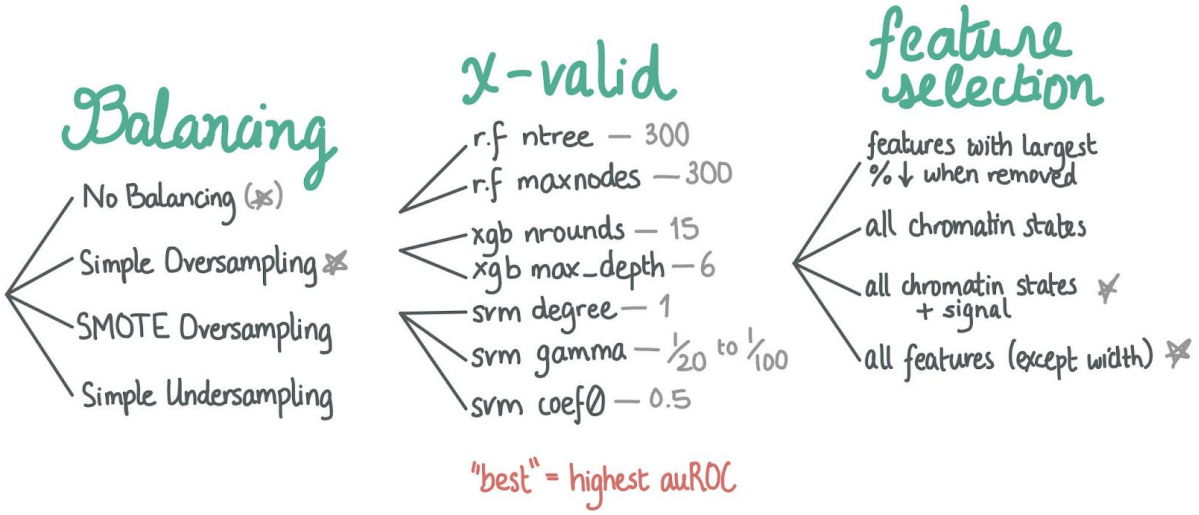

Figure 2: steps for choosing best model which were done by automation; sequence of choices from left to right

Applications. Incidentally, the ML classifier could also be applied to doing single-cell level research. Single-cell detail can give insights on cell-type and conditional specificity of DNA regulation, yet getting data at this level is difficult, especially for ChIP-seq. Using the above-mentioned classifier, pseudo-bulk ATAC data could be for example used to predict single-cell level H3K27ac ChIP peaks.

Are there properties which distinguish the overlapping (H3K27ac ChIP and ATAC) peaks and the unique peaks (either type)?

Looking into the features supplied to the classifier, some significant differences could be found between overlapping and unique peaks (as expected from the classifier's prediction power). A comprehensive table with all features and comparisons between overlapping and unique peaks can be found in Appendix 2. As an example, signal (from bigWig file) of the ATAC or ChIP peak summit (summit found from narrowPeaks file) is significantly higher in overlaps than unique peaks (Figure 3). GC content is also significantly higher for overlapping regions of H3K27ac ChIP and ATAC, rather than unique regions. This is consistent with the prevalence of CpG islands around promoters, indicating more H3K27 acetylation around promoter regions.

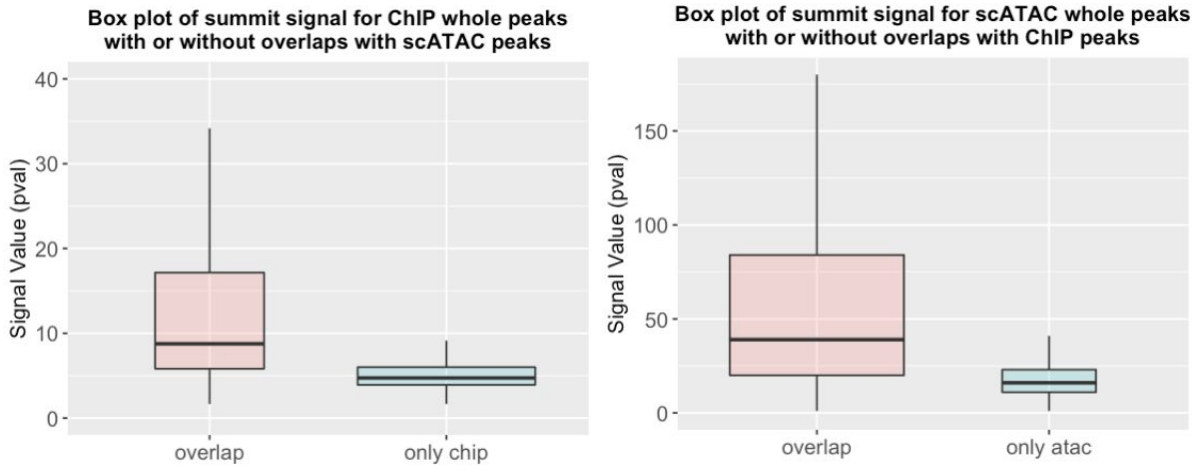

Figure 3: Box plots[8] for summit signal comparison between overlapping and non-overlapping (unique) peaks. Left: H3K27ac ChIP peaks set, right: ATAC. Table with other features included in Appendix 2.

[8] Width of bars corresponds to relative sample size.

Furthermore, looking at the distances of peaks to transcription start sites (TSS) and enhancers (from GENCODE and the EnhancerAtlas, respectively), more unique ATAC peaks lie far from TSS, while most H3K27ac ChIP peaks lie close to TSS, showing that ChIP peaks lie around promoters, and further from enhancers.

A final comment is on chromatin states: for each peak, the proportion of the peak range in a particular chromatin state was calculated (Figure 4), and average across peaks (separated by overlap or unique) was taken for the table in Appendix 2. Although during comparison, some of these were the least "different" (highest p-value), in the ML classifier, using chromatin state information was deemed important during feature selection. This probably arises because the chromatin states individually may not give much importance, yet using all chromatin state data together for the classifier may be powerful because it integrates many factors; it links both types of distances (distance to TSS and to enhancers), and additionally incorporates implicit factors like transcription factor binding, and locations of other histone modifications. A decision tree-like classifier could be built from this information, hence probably the suitability of the random forest model for ATAC data, and the logistic regression model for H3K27ac ChIP.

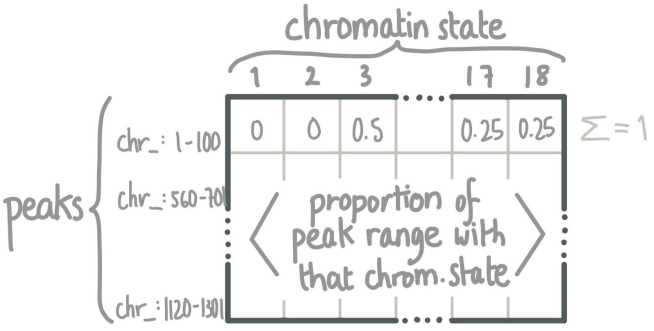

Figure 4: illustration of how chromatin state information was used

Can we find possible AD GWAS SNPs from accessibility data?

Only preliminary results were found for this question.

Background on gkmSVM. Gapped k-mer SVM can predict similarity of sequences based on k-mers with gaps of mismatches. Corces *et al*.[9] used this method to predict which sequences were accessible by pseudo-bulk ATAC and not, and consequently testing whether sequences around a GWAS Alzheimer's Disease (AD) SNP effect allele and non-effect allele are differentially accessible. Their data could even show which SNPs were potentially functional, and identify affected transcription factor (TF) binding motifs.

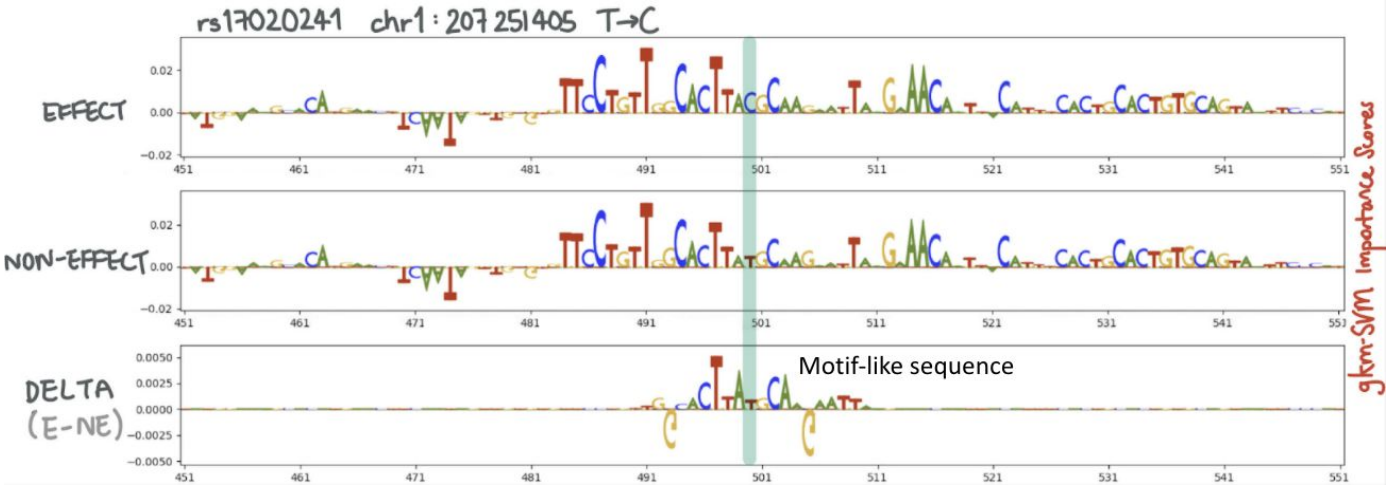

Figure 5: graphs of gkmSVM Importance Scores for each nucleotide for an example GWAS SNP (located at position 500, nucleotide highlighted in green)

[9] "Single-cell epigenomic identification of inherited risk loci in Alzheimer's and Parkinson's disease", Corces *et al.*, 2020

What has been accomplished. After training the gkmSVM model on our pseudo-bulk ATAC data (replicating what was done in the Corces *et al.*[9] paper), graphs like in Figure 5 were produced, which show "gkmSVM Importance Scores" (their pipeline output) for each nucleotide surrounding each SNP. When the difference was calculated between the Importance Scores of sequences with the effect versus non-effect alleles, a delta value was found for each nucleotide, and a motif-like sequence deemed "important" by gkmSVM emerged (Figure 5). However, these results did not seem to be very clear; more refinement and a better understanding would be required. Moreover, when a preliminary analysis on ATAC-QTL correlation with the significant GWAS SNPs found by gkmSVM was done, no interesting correlation was found (Appendix 3).

What we would like to do. Going forward, we can train gkmSVM models on either ATAC or H3K27ac ChIP data or other combinations (e.g. peaks of ATAC and ChIP overlap), and in addition better integrate the predictions with ATAC-QTL SNP data, for example, to more fully understand AD GWAS SNPs.

Conclusive remarks

We can conclude that there are indeed significant differences between ATAC and H3K27ac ChIP overlapping peaks versus peaks unique to ATAC or ChIP. A machine learning classifier can be trained to predict these differences. Some important features studied were peak (bigWig) signal, GC content, distances from promoter and enhancer, and chromatin states. The classifier can potentially be used for finding which pseudo-bulk ATAC peaks overlap with H3K27ac ChIP when single cell-level ChIP data is unavailable.

Finally, a gapped k-mer SVM algorithm was explored, which could possibly show the positions of differentially accessible locations. This would be especially interesting in areas around SNPs to understand disease-causing loci. Previous research has also found that it is possible to identify important transcription factor binding motifs using gkmSVM as well.

For future development, it may be of interest to investigate time sequence of ATAC accessibility and H3K27ac modifications by pseudotime analysis; there seems to be a much higher proportion of non-overlapping (unique) H3K27ac ChIP peaks than non-overlapping ATAC peaks (Figure 6). This seems counter-intuitive as the signal for "active promoters and enhancers" (H3K27 acetylation) should be at areas of open chromatin (ATAC accessible). Could this mean that maybe there is a time element, where the areas of H3K27 acetylated chromatin is *later* made accessible?

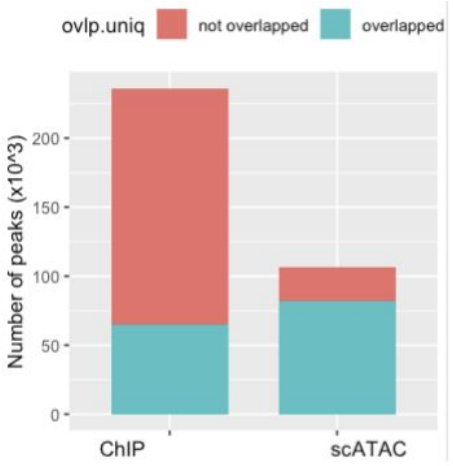

Figure 6: stacked bar graph showing absolute numbers of peaks, separated into those overlapping or not with the other type of peak

Acknowledgements

I am a million times grateful to Lei Hou and Xushen Xiong who have guided me throughout this project, and for the many thoughtful discussions we have had. I am also thankful for the Broad Institute for providing the server, resources, and help needed to work with these large datasets. Finally, I would like to thank Manolis Kellis for allowing me to partake in this internship, and to the whole Kellis lab for the intriguing weekly lab group meetings.

Works Cited

1. **'An integrated encyclopedia of DNA elements in the human genome'**, ENCODE project consortium, 2012
2. **'Enhanced Regulatory Sequence Prediction Using Gapped k-mer Features'**, Ghandi *et al.*, 2014, PLOS Comp Bio
3. **'Integrative analysis of 111 reference human epigenomes'**, Roadmap Epigenomics Consortium *et* al., 2015
4. **'Genetic and epigenetic fine mapping of causal autoimmune disease variants'**, Farh *et al.*, 2015
5. **'gkmSVM: an R package for gapped-kmer SVM'**, Ghandi *et al.*, 2016
6. **'A systematic study of the class imbalance problem in convolutional neural networks'**, Mazurowski *et al.*, 2017
7. **'A histone acetylome-wide association study of Alzheimer's disease identifies disease-associated H3K27ac differences in the entorhinal cortex'**, Marzi *et al.*, 2018
8. **'The Post-GWAS Era: From Association to Function'**, Gallagher & Chen-Plotkin, 2018
9. **'Integrating ChIP-seq with other functional genomics data'**, Shan Jiang & Ali Mortazavi, 2018
10. **'A review on handling imbalanced data'**, Spelmen & Porkodi, 2018
11. **'Activity-by-contact model of enhancer-promoter regulation from thousands of CRISPR perturbations'**, *Broad Institute*, 2019
12. **'Brain cell type–specific enhancer–promoter interactome maps and disease-risk association'**, Nott *et al.*, 2020
13. **'Single-cell epigenomic identification of inherited risk loci in Alzheimer's and Parkinson's disease'**, Corces *et al.*, 2020
14. **'Single cell epigenomic atlas of the developing human brain and organoids'**, Ziffra *et al.*, 2020

15. **https://galaxyproject.github.io/training-material/topics/epigenetics/tutorials/atac-seq/tutorial.html**
16. **https://github.com/ENCODE-DCC/chip-seq-pipeline2**
17. **https://www.archrproject.com/bookdown/index.html**

Appendix 1: ML classifier results

Difference between each image. The following images show tables with different metrics (auROC, F1 score, and auPRC) of the ML classifier performance. This was done for classifiers trained and tested on different samples of peaks (chromosomes 21, 22, X, and a random sample of 5000 genome wide peaks). Training on larger sets were computationally too slow; an improvement would be to increase efficiency of the code, or to split the task, so as to be able to train the classifier using more example peaks. However, it is interesting to see that there are differences in performance between the different chromosomes, showing heterogeneity among the peaks in the different chromosomes.

Top and bottom tables. The top table in each image is for the classifier trained to discern H3K27ac ChIP peaks which overlap with ATAC or not, and the bottom table is a corresponding one for ATAC peaks.

Explanation of columns. The coloured columns show the "improvement" of the classifier's metrics when tested on the validation dataset at each step of the sequential improvement method (as illustrated in Figure 2). The following is an explanation of each column:

- baseline_2.features: use just 2 features (peak widths, and summit (bigWig) signal)
- baseline_all.features: use all features
- balance_data: use all features, plus balance the positive and negative sets (best balancing technique is chosen)

- cross.validation_hyperparam: use all features, best balancing technique, and best hyperparameters for each respective model
- feature.selection: use best balancing technique, best hyperparameters for each respective model, and choose the best set of features.

The final column, "test.data" is the performance of the final model (the model after "feature.selection") used on the test dataset (instead of validation set as used in the middle columns).

Chromosome 21

Using ChIP to predict overlaps with scATAC

Percentages		baseline_2.features	baseline_all.features	balance_data	cross.validation_hyperparam	feature.selection	test.data
auROC baseline =	logistic	82.67%	88.04%	88.14%	88.14%	87.92%	86.33%
	random.forest	79.55%	88.79%	89.14%	89.12%	87.49%	86.59%
	xgboost	81.89%	88.53%	88.53%	88.81%	86.28%	84.41%
50%	svm	83.05%	86.05%	87.21%	88.98%	87.98%	86.66%
F1 Score baseline =	logistic	63.31%	70.69%	70.75%	70.75%	71.39%	69.06%
	random.forest	63.10%	71.98%	71.66%	72.04%	69.81%	68.14%
	xgboost	65.86%	71.11%	71.11%	70.30%	67.47%	68.25%
35%	svm	56.35%	70.27%	72.44%	73.48%	71.59%	69.92%
auPRC baseline =	logistic	90.10%	93.88%	93.85%	93.85%	93.83%	93.26%
	random.forest	79.93%	93.99%	94.33%	93.87%	93.26%	93.67%
	xgboost	89.11%	93.59%	93.59%	94.36%	92.92%	91.85%
73%	svm	89.55%	91.86%	93.26%	94.14%	93.88%	93.52%

Using scATAC to predict overlaps with ChIP

Percentages		baseline_2.features	baseline_all.features	balance_data	cross.validation_hyperparam	feature.selection	test.data
auROC baseline =	logistic	82.66%	86.51%	86.57%	86.57%	86.71%	86.19%
	random.forest	78.14%	88.95%	89.12%	89.40%	89.92%	89.81%
	xgboost	79.82%	87.35%	87.35%	88.26%	88.87%	87.55%
50%	svm	74.26%	79.53%	81.84%	84.96%	87.39%	88.41%
F1 Score baseline =	logistic	88.66%	91.50%	91.47%	91.47%	91.30%	89.24%
	random.forest	88.21%	91.65%	91.88%	91.97%	91.31%	90.66%
	xgboost	88.62%	90.28%	90.28%	91.52%	91.38%	90.51%
61%	svm	87.90%	88.48%	89.40%	90.06%	90.20%	90.05%
auPRC baseline =	logistic	49.28%	59.17%	59.93%	59.93%	60.06%	56.98%
	random.forest	40.62%	67.02%	67.17%	67.66%	65.71%	66.36%
	xgboost	43.45%	61.09%	61.09%	64.14%	63.81%	61.22%
21%	svm	35.18%	46.45%	55.99%	56.59%	59.38%	61.35%

Chromosome 22

Using ChIP to predict overlaps with scATAC

Percentages		baseline_2.features	baseline_all.features	balance_data	cross.validation_hyperparam	feature.selection	test.data
auROC baseline =	logistic	86.83%	92.21%	92.21%	92.21%	91.43%	91.53%
	random.forest	84.14%	92.94%	92.90%	92.50%	91.83%	90.46%
	xgboost	86.08%	91.86%	92.31%	91.76%	91.03%	91.12%
50%	svm	86.18%	88.38%	88.88%	92.06%	91.27%	91.37%
F1 Score baseline =	logistic	70.85%	74.46%	74.46%	74.46%	73.99%	76.11%
	random.forest	65.33%	75.77%	76.07%	75.69%	73.70%	77.49%
	xgboost	68.46%	75.11%	74.61%	75.16%	71.56%	75.05%
33%	svm	67.82%	74.81%	75.00%	74.16%	74.35%	75.06%
auPRC baseline =	logistic	93.31%	96.79%	96.79%	96.79%	96.43%	96.09%
	random.forest	74.94%	96.53%	96.93%	84.40%	84.76%	86.95%
	xgboost	90.54%	92.21%	95.21%	91.51%	92.81%	92.33%
75%	svm	93.09%	93.36%	92.95%	96.80%	96.48%	96.05%

Using scATAC to predict overlaps with ChIP

Percentages		baseline_2.features	baseline_all.features	balance_data	cross.validation_hyperparam	feature.selection	test.data
auROC baseline =	logistic	81.96%	86.02%	87.02%	87.02%	86.79%	88.07%
	random.forest	72.90%	82.61%	84.38%	83.93%	87.14%	84.11%
	xgboost	77.40%	82.32%	83.98%	87.18%	87.05%	86.32%
50%	svm	62.52%	74.72%	86.44%	87.70%	87.34%	86.64%
F1 Score baseline =	logistic	92.67%	93.08%	93.18%	93.18%	93.10%	93.54%
	random.forest	92.78%	93.11%	93.06%	92.74%	93.61%	92.40%
	xgboost	92.67%	93.08%	93.22%	93.63%	93.49%	92.90%
63%	svm	92.51%	92.96%	93.65%	93.57%	93.13%	93.13%
auPRC baseline =	logistic	34.49%	42.19%	43.87%	43.87%	44.11%	48.48%
	random.forest	24.36%	39.13%	38.42%	34.85%	49.16%	33.54%
	xgboost	25.71%	38.42%	40.65%	48.56%	48.23%	39.79%
13%	svm	15.59%	29.47%	47.02%	44.97%	43.89%	42.65%

Chromosome X

Using ChIP to predict overlaps with scATAC

Percentages		baseline_2.features	baseline_all.features	balance_data	cross.validation_hyperparam	feature.selection	test.data
auROC baseline =	logistic	81.38%	89.13%	89.48%	89.48%	89.67%	85.42%
	random.forest	79.23%	91.42%	91.40%	91.01%	**90.56%**	**87.88%**
	xgboost	80.79%	89.95%	90.24%	90.14%	**90.14%**	**87.72%**
50%	svm	79.59%	89.00%	89.00%	89.84%	89.34%	84.99%
F1 Score baseline =	logistic	70.91%	78.37%	79.06%	79.06%	**80.55%**	73.29%
	random.forest	65.83%	80.32%	80.44%	80.70%	78.98%	**76.55%**
	xgboost	69.03%	79.15%	78.53%	78.71%	77.65%	**76.06%**
43%	svm	68.61%	80.52%	80.45%	80.20%	79.11%	74.32%
auPRC baseline =	logistic	82.50%	88.52%	88.79%	88.79%	88.92%	86.65%
	random.forest	76.40%	93.44%	93.10%	92.25%	91.04%	88.00%
	xgboost	83.74%	91.63%	92.45%	91.75%	**92.03%**	**90.07%**
62%	svm	80.29%	89.51%	89.51%	90.05%	88.65%	85.45%

Using scATAC to predict overlaps with ChIP

Percentages		baseline_2.features	baseline_all.features	balance_data	cross.validation_hyperparam	feature.selection	test.data
auROC baseline =	logistic	78.87%	84.04%	84.04%	84.04%	84.62%	84.11%
	random.forest	77.67%	85.74%	85.63%	86.05%	**86.18%**	**86.28%**
	xgboost	79.78%	83.85%	83.85%	85.01%	85.70%	85.97%
50%	svm	76.26%	79.55%	81.26%	84.12%	84.57%	84.36%
F1 Score baseline =	logistic	79.52%	80.78%	80.78%	80.78%	81.90%	81.65%
	random.forest	78.40%	84.41%	84.03%	84.50%	**84.21%**	**84.15%**
	xgboost	78.23%	82.18%	82.18%	82.85%	83.73%	83.19%
55%	svm	77.73%	79.54%	80.92%	80.74%	81.91%	83.27%
auPRC baseline =	logistic	68.45%	73.50%	73.50%	73.50%	74.55%	74.98%
	random.forest	65.58%	76.32%	76.35%	76.86%	**76.92%**	**76.24%**
	xgboost	66.42%	73.94%	73.94%	74.85%	75.00%	75.71%
40%	svm	64.74%	66.73%	69.32%	74.38%	74.87%	75.19%

Whole Genome Random 5000 peaks

Using ChIP to predict overlaps with scATAC

Percentages		baseline_2.features	baseline_all.features	balance_data	cross.validation_hyperparam	feature.selection	test.data
auROC baseline =	logistic	83.64%	89.46%	89.46%	89.46%	**88.99%**	**88.64%**
	random.forest	79.40%	89.58%	89.67%	89.38%	88.78%	88.42%
	xgboost	82.94%	89.44%	89.44%	89.64%	**89.47%**	**88.57%**
50%	svm	83.22%	87.21%	87.21%	89.21%	88.57%	88.02%
F1 Score baseline =	logistic	77.28%	82.08%	82.08%	82.08%	81.40%	**81.05%**
	random.forest	72.83%	82.82%	82.80%	82.94%	**82.23%**	80.78%
	xgboost	76.09%	81.93%	81.93%	82.34%	**81.80%**	**81.40%**
50%	svm	77.25%	81.31%	81.28%	81.91%	81.65%	80.82%
auPRC baseline =	logistic	78.38%	87.34%	87.34%	87.34%	**86.65%**	**87.04%**
	random.forest	74.45%	87.00%	86.93%	85.62%	85.34%	84.28%
	xgboost	76.73%	86.74%	86.74%	87.40%	**87.29%**	86.48%
50%	svm	78.04%	84.49%	84.49%	86.91%	86.09%	86.07%

Using scATAC to predict overlaps with ChIP

Percentages		baseline_2.features	baseline_all.features	balance_data	cross.validation_hyperparam	feature.selection	test.data
auROC baseline =	logistic	83.24%	87.03%	87.06%	87.06%	86.98%	83.57%
	random.forest	80.11%	87.35%	87.43%	87.34%	**87.14%**	83.11%
	xgboost	82.44%	87.05%	87.05%	87.26%	**87.25%**	**83.83%**
50%	svm	81.44%	83.83%	84.92%	87.12%	87.01%	**83.57%**
F1 Score baseline =	logistic	75.82%	80.08%	80.24%	80.24%	**80.31%**	76.12%
	random.forest	73.77%	79.92%	79.64%	79.70%	79.44%	76.08%
	xgboost	75.25%	79.25%	79.25%	79.55%	79.64%	**76.89%**
50%	svm	73.68%	77.16%	77.72%	80.30%	**80.40%**	76.36%
auPRC baseline =	logistic	79.54%	83.21%	83.26%	83.26%	83.16%	**79.59%**
	random.forest	71.27%	84.90%	84.83%	84.83%	**84.73%**	77.25%
	xgboost	78.38%	84.48%	84.48%	85.02%	**85.06%**	**79.95%**
50%	svm	77.12%	79.11%	79.89%	83.67%	83.48%	**79.54%**

Appendix 2: Features comparison between overlap and unique peaks

Features	H3K27ac ChIP peaks			scATAC peaks		
	Overlapping scATAC	Uniquely ChIP	p-value	Overlapping H3K27ac ChIP	Uniquely scATAC	p-value
n_peak.width	3107.47455	719.0795495	0	610.1187325	365.0002099	0
n_GC.content	0.480454812	0.4610040551	0	0.5149705273	0.4505029985	0
n_CG.dinuc.freq	2.046808711	1.189881976	0	2.861142278	1.279192886	0
n_summit.signal	15.09738478	5.263948074	0	64.53624739	19.85674896	0
n_mean.peak.signal	6.589543861	3.313265229	0	47.398336	17.33579913	0
i_dist.range.geneTSS	21094.15177	26741.90663	2.15E-290	20946.07516	34420.6529	0
i_dist.summit.geneTSS	21998.56618	27086.99932	4.38E-233	21147.15436	34597.15575	0
i_dist.range.gene	4278.004166	3672.262357	1.15E-17	3862.164047	10910.45123	4.20E-288
i_dist.range.txptTSS	12517.29837	14957.06978	1.40E-120	11964.71154	24478.3539	0
i_dist.summit.txptTSS	13185.26574	15280.52002	6.24E-88	12143.84067	24651.64136	0
i_dist.range.txpt	4279.143704	3673.991546	1.24E-17	3863.054711	10917.05641	1.52E-288
dist.range.enhancer	50493.03117	72374.53047	2.64E-173	47437.31893	107546.4505	8.86E-266
n_chrom.state.1	0.02981814372	9.35E-05	0	0.06875496324	0.001738565206	0
n_chrom.state.2	0.01571242394	0.0004338591945	0	0.03157690539	0.003977996497	0
n_chrom.state.3	0.02506305098	0.0008103003608	0	0.05182078271	0.00383547018	0
n_chrom.state.4	0.0266862915	0.001380404006	0	0.03238837942	0.006026694412	0
n_chrom.state.5	0.04452754022	0.1560546999	0	0.03643758171	0.02415153738	9.29E-26
n_chrom.state.6	0.0844771545	0.1890145098	0	0.05831614801	0.05777430136	0.7395356439
n_chrom.state.7	0.01325836399	0.012152134	0.005810668099	0.01266855616	0.005812574737	1.96E-31
n_chrom.state.8	0.005821151466	0.002469194105	4.06E-46	0.007122675235	0.000652347262	1.03E-95
n_chrom.state.9	0.08781379366	0.01586731799	0	0.1366053406	0.05325117046	0
n_chrom.state.10	0.002836544206	0.00125224967	2.75E-32	0.004148865627	0.002579751482	1.31E-06
n_chrom.state.11	0.1946186364	0.08154456623	0	0.2178009368	0.2638857505	2.43E-53
n_chrom.state.12	0.006247986665	0.012167571	7.77E-65	0.004721617813	0.01712577919	2.36E-49
n_chrom.state.13	0.0148380236	0.03170866456	2.86E-176	0.01247023262	0.04592811538	8.34E-128
n_chrom.state.14	0.01439279041	0.001016332874	0	0.02263496781	0.003274778197	1.88E-267
n_chrom.state.15	0.02646574441	0.006506240663	0	0.03292745985	0.01675948137	1.15E-64
n_chrom.state.16	0.03151537861	0.03347816365	0.004017557276	0.03045477098	0.03094814601	0.6852245567
n_chrom.state.17	0.0442695836	0.06929980066	1.17E-166	0.03583721058	0.06714687433	1.77E-75
n_chrom.state.18	0.2112547934	0.3794439555	0	0.1761036106	0.3918717684	0

Columns. On the left side of the double red lines is data for H3K27acChIP peaks, and on the right is data for ATAC peaks. For each, there are 3 columns: average value of the feature for the peaks overlapping the other type of peak (ChIP or ATAC), that of the unique peaks (not overlapping), and the p-value from a student's t-test comparing the overlapping to unique peaks.

Rows. These are the various features which were studied, including peak width (size), GC content, CG dinucleotide frequency, bigWig signal at summit or across the entire peak, distance to transcription start sites (TSS) or genes (several different ways of measuring were used, to see which one worked the best), distance to enhancer, and the proportion of peaks in different chromatin states.

Appendix 3: gkmSVM SNPs and ATAC-QTL

Credits for this analysis go to Xushen Xiong.

There seems to be no correlation between ATAC-QTLs and gkmSVM-predicted important GWAS SNPs. These were ranked from the largest sum(abs(*delta values for all nucleotides around the SNP*)) - Figure 5 can help visualise this.