The Dynamic Relationship between Gene Expression and Epigenetically Modified Enhancers

YUN-HUA (ESTHER) HSIAO and HUBERT JENQ, University of California, Los Angeles

General Terms: Gene Expression

Additional Key Words and Phrases: Liquid Association, histone modification, enhancer, gene expression, epigenome

ACM Reference Format:

Yun-Hua (Esther) Hsiao and Hubert Jeng. 2015. The Dynamic Relationship between Gene Expression and Epigenetically Modified Enhancers. *ACM Trans. Appl. Percept.* 1, 1, Article 1 (June 2015), 9 pages.
DOI:http://dx.doi.org/10.1145/0000000.0000000

1. INTRODUCTION

Gene expression is tightly regulated by various means including proximal and distal DNA sequence elements usually in *cis* with the target genes [Fernandez 2012]. The promoter region, located at the 5′ end of the gene, is the proximal *cis*-regulatory element where the transcription factors assemble. Other *cis*-regulatory elements such as enhancers and silencers are located in either introns or exons of the target genes, or in distant sites. Enhancers contain short DNA sequence motifs for different transcription factors to bind and activate transcription [Shlyueva 2014]. Genome-wide chromatin epigenetic maps also showed that enhancers display the largest variability in their activation states across diverse cell types [Heintzman 2009], which suggests that enhancers play important roles in the development and differentiation of different cell types by activating cell type-specific gene expression programs. Experimental procedures like ChIP-Seq is a popular method to investigate the post-translational modification of the histones which includes phosphorylation, acetylation and methylation. Different histone modifications can mediate gene expression changes by activating or inhibiting transcription.

In this study, we investigated the dynamic regulation of gene expression by enhancers with epigenetic marks using the imputed data from Jason Ernst [Ernst 2015]. In order to do this, we adopted the Generalized Liquid Association (GLA) method originally developed by Ker-Chau Li [Li 2002], and later updated by Yen-Yi Ho [Ho 2009], to identify: (1) co-expressed gene pairs mediated by an enhancer with epigenetic mark, and (2) genes whose expression is modulated by enhancer pairs with epigenetic marks. Ho *et al.* improved on the

Author's address: Y.H.E. Hsiao, 610 Charles E. Young Drive East, Terasaki Life Sciences Building, Room 2000B, Los Angeles, CA 90095-1570; email: esther.yh.hsiao@ucla.edu; H. Jenq, hjenq@ucla.edu.

DOI:http://dx.doi.org/10.1145/0000000.0000000

(a) Question 1 (b) Question 2

Fig. 1: General analysis pipelines. We downloaded 16 experimentally validated human enhancer locations in chromosome 21 from the VISTA Enhancer Browser. We overlapped the enhancer regions with the provided imputed histone mark signals, and used the average signal normalized by regions overlapped in the enhancer regions in the GLA calculation. In Q1, we first filtered out the genes that are known to correlate well in the gene expression. We only collected the gene pairs with gene expression correlation that has Pearson's correlation coefficient between -0.1 and 0.1 for GLA calculation. For both Q1 and Q2, we used the GLA R package to calculate GLA scores and estimate the P-values.

previously derived Liquid Association (LA) in the cases where the conditional mean or variances depend on Z. In this scenario, the $E[XYZ]$ can be a non-zero value despite the correlation of X and Y conditioned on Z is constant. In GLA, the $g(x)$ function becomes the correlation of X and Y conditioned on Z is instead of the expected value of X and Y conditioned on Z. As a result, the GLA statistics is not affected by conditional variances or means that depend on Z.

2. RESULTS

2.1 Identification of significant GLA triplets

To investigate the role of enhancers with epigenetic markers in modulating the gene expression, we asked the following two questions: (Q1) What enhancers could mediate the co-expression of gene pairs? (Q2) What genes are under the regulation of epigenetically modified enhancers? In answering Q1, we first obtained a list of 52 genes whose gene expression are not correlated (Pearson's correlation coefficient between -0.1 and 0.1) on human chromosome 21 based on GeneFriends database [van Dam et al. 2014], and formed all possible gene pairs as X and Y in the GLA setup. We obtained all the known enhancers in human chromosome 21 and used the epigenetic mark signals in the enhancer regions as Z for GLA calculation (Figure 1).

Using the GLA R package developed by [Ho 2009], we obtained 31,290 triplets (0.4%) out of total 7,289,984 triplets and 31,367 triplets (1%) out of 2,747,264 total triplets that have significant GLA scores with both P-value < 0.01 and absolute GLA scores > 0.2 for Q1 and Q2 respectively. We spot checked the correlation scatter plots of one triplet with positive GLA score and one with negative GLA score in Q1 and Q2 separately (Figure 2) to confirm a true LA relationship. As shown in Figures 2A and 2B, epigenetically modified enhancers could dynamically correlate with the co-expression and counter-expression of two genes. Likewise

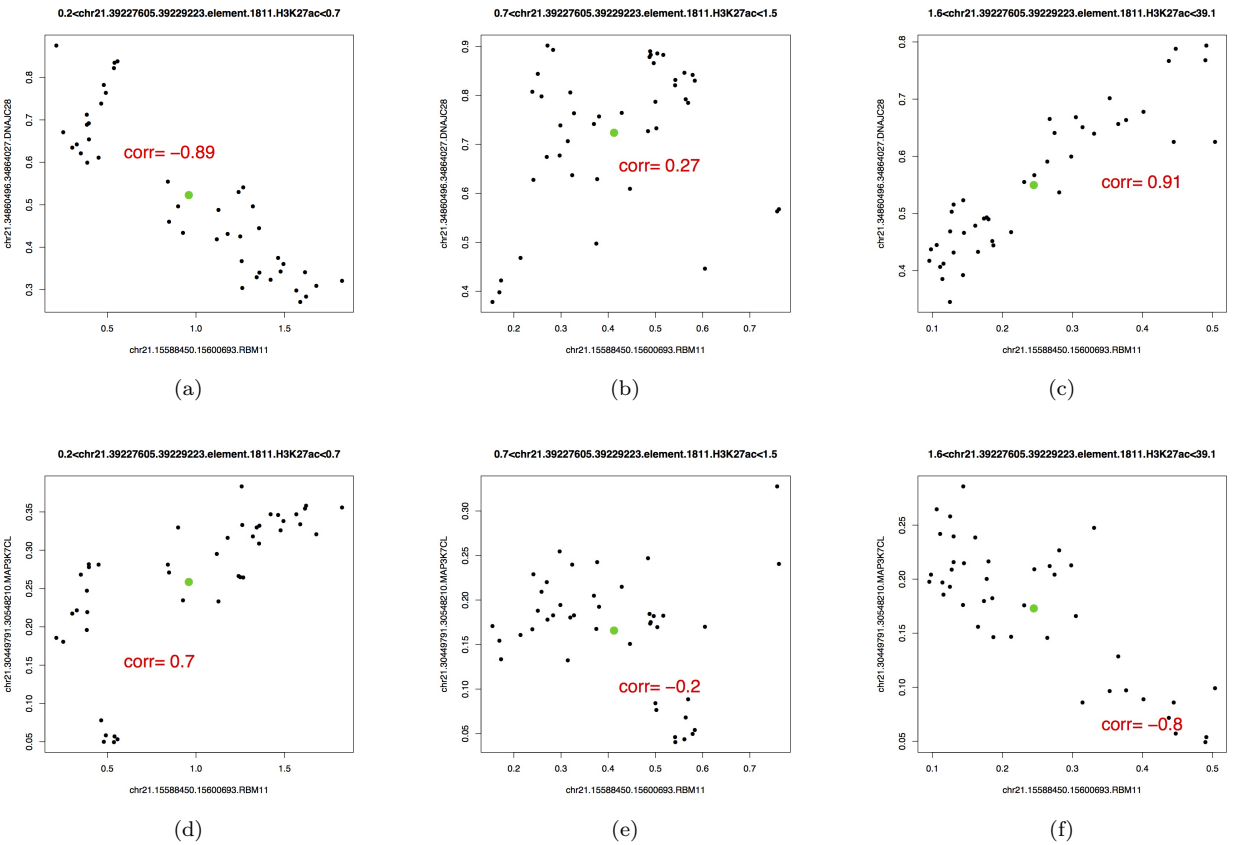
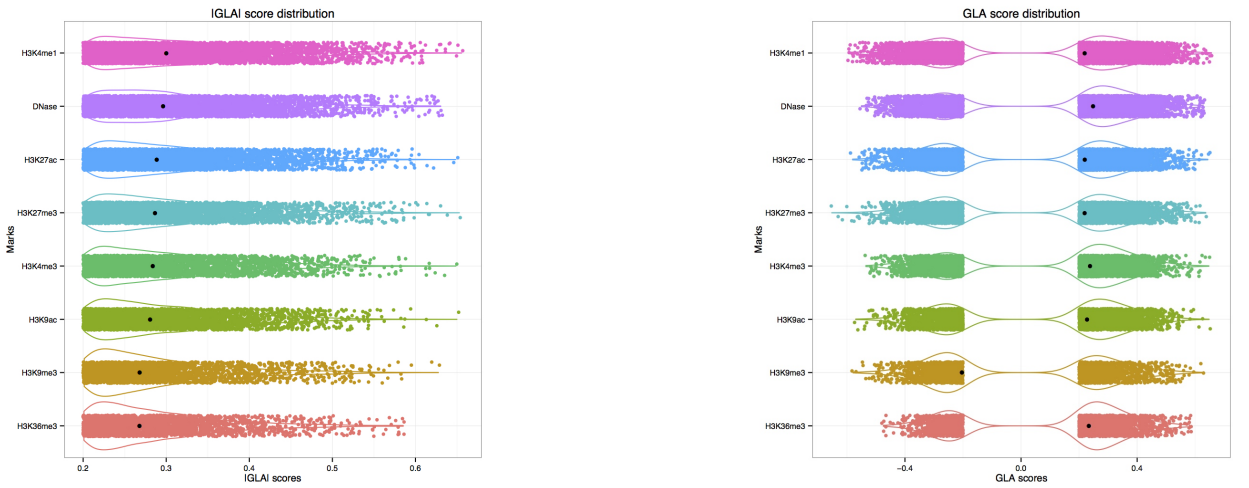


Fig. 2: Example GLA scatter plots. These plots were generated by the GLA R package. From left to right, each plot show the correlation between X and Y conditioning on different values of Z. The program tries to make the number of data points in each Z range similar. The green dots represent the median of X and Y. (a,b,c) Q1 example with GLA > 0. (d,e,f) Q1 example with GLA < 0.

in Figures 2C and 2D, gene expression of a gene could relate to the correlation of various enhancers with different epigenetic marks dynamically.

To examine whether there is a specific histone mark and/or enhancer that have interesting patterns in mediating the co- or counter-expression of paired genes, we compared the GLA score distributions as well as the distributions of GLA score magnitudes across histone marks and enhancers (Figure 3). Among all histone marks, H3K4me1 has the highest |GLA scores|. This result is consistent with previous global studies (1-7) that showed H3K4me1 has significant enrichment near the transcription start sites (TSSs) and is essential in gene activation. H3K36me3 has the lowest |GLA scores|, but the vast majority of the significant triplets involving H3K36me3 have positive GLA (median > 0). This observation is in line with previous findings (1-7) that H3K36me3 is associated with gene activation, which also suggests that high level of H3K36me3 tend to mediate the co-expression of the corresponding gene pairs in a small degree. On the other hand, H3K9me3, known to associated with transcriptional repression, has more negative GLA scores (median < 0). This H3K9me3 result is reasonable because higher mark signal is associated with the negative correlation of



(a) Magnitude of GLA score

(b) GLA score

Fig. 3: Significant GLA score distribution for each histone mark. The violin plots show (A) the magnitude of significant GLA scores and (B) the overall GLA scores for each histone mark. The violin plots are overlapped with jitters showing the exact value for each significant triplet. The black dots represent the mean values of the distributions.

the gene expression between gene pairs, implying one of the genes in the pair can have lower gene expression level.

2.2 The enhancers are significantly closer to the associated genes in the corresponding significant triplets

We first compared the distance metric between all significant GLA score triplets and random triplets. The significant GLA triplets where X and Y are enhancers and Z is a gene (Q2) are significantly closer together in 1-D distance than the random triplets. We then examined the distance metric for only one type mark. When X and Y are genes and Z is a enhancer (Q1), looking only at H3K27me3 enhancer scores showed that the significant GLA triplets were significantly closer than the random triplets as shown in figure 4.

In Q2, H3K27me3 and H3K4me1 significant GLA triplets were significantly closer than random triplets which can be seen in table 1. Through this analysis, we see that some significant GLA triplet pairs are closer in linear distance to each other indicating that proximity may be a method of their Liquid Association interactions.

2.3 Significant triplets involving enhancers with H3K27me3 has higher interaction with corresponding genes

To follow up the distance analysis, we used Hi-C data at 40kb resolution for human embryonic stem cells (ES) and human fibroblasts (IMR90). To calculate the interaction score, we looked at the position overlap between $X - Z$ and $Y - Z$ and summed their corresponding normalized interaction score.

In both Q1 and Q2 for all mark, enhancer, and gene combinations, the significant GLA triplets were not found to significantly interact more with each other compared to random triplets. However, in Q1, when looking at only H3K27me3 and H3K9me3, the significant GLA triplets were found to interact more than random triplets in the ES Hi-C data as evident in figure 5.

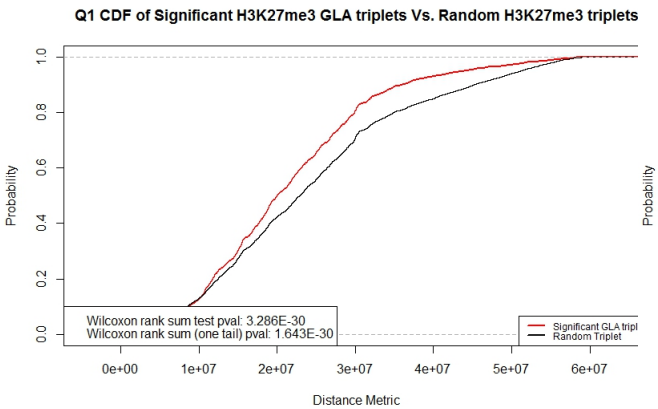


Fig. 4: Q1 Distance CDF plot for only H3k27me3 enhancer-marks. We see that the significant GLA triplet CDF line is above the random triplets indicating that the significant GLA triplets are closer together on average. The corresponding one-tail wilcoxon rank sum p-value is 1.643E-30.

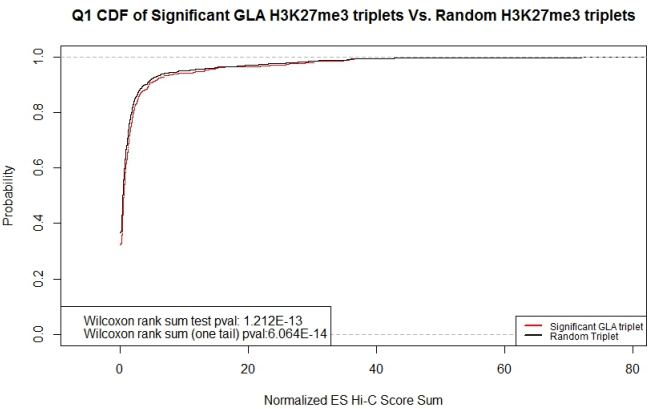


Fig. 5: Q1 Interaction score CDF plot for only H3k27me3 enhancer-marks. The plot shows that the significant GLA triplet curve is below the random triplet curve, indicating that significant GLA triplets have a higher normalized ES Hi-C score sum. This is reflected in the one-tail wilcoxon rank sum p-value of 6.064E-14.

Q1:GLA(gene X, gene Y|enhancer Z)						
Histone Marks	**Dist/2-tail**	**Dist/1-tail**	**Hi-C.ES/2-tail**	**Hi-C.ES/1-tail**	**Hi-C.IM/2-tail**	**Hi-C.IM/1-tail**
H3K27ac	4.06E-06	0.999997971	0.374296051	0.812852576	0.991008774	0.495504387
H3K27me3	3.29E-30	1.64E-30	1.21E-13	6.06E-14	0.000165704	8.29E-05
H3K36me3	2.07E-17	1	0.000385165	0.99980742	6.66E-07	0.999999667
H3K4me1	0.152755744	0.923622407	0.629120994	0.685440211	0.432463661	0.21623183
H3K4me3	0.020121136	0.9899395	0.781866785	0.609067609	0.500282724	0.74985945
H3K9ac	1.06E-06	0.99999947	0.287196311	0.856402449	0.507244337	0.253622168
H3K9me3	0.002512867	0.001256433	4.32E-07	2.16E-07	0.425937327	0.212968664
Q2:GLA(enhancer X, enhancer Y|gene Z)						
Histone Marks	**Dist/2-tail**	**Dist/1-tail**	**Hi-C.ES/2-tail**	**Hi-C.ES/1-tail**	**Hi-C.IM/2-tail**	**Hi-C.IM/1-tail**
H3K27ac	0.872946792	0.563531561	0.87689328	0.561560216	0.561219391	0.280609695
H3K27me3	8.00E-08	4.00E-08	0.026772447	0.986614595	0.01979045	0.990105358
H3K36me3	0.916075675	0.458037837	0.277784791	0.861120156	0.797991416	0.601024334
H3K4me1	0.033424576	0.016712288	0.090528702	0.045264351	0.215729329	0.107864664
H3K4me3	0.936982785	0.468491392	0.413298255	0.793357241	0.528486778	0.73576328
H3K9ac	0.705293003	0.647360198	0.409742769	0.204871384	0.450868298	0.225434149
H3K9me3	7.54E-08	3.77E-08	0.920337536	0.539842714	0.565633508	0.282816754

P-values were calculated by Wilcoxon rank sum test. Dist: Distance analysis. Hi-C.tissue: Hi-C analysis using the interaction map in the corresponding tissue.

Table I. : Distance and Interaction analysis *P*-value summary.

In the IMR90 data, the H3K27me3 significant GLA triplets were found to interact more than the random triplets as seen in table 1. From this, we can extend our hypothesis that certain liquid associations found may exhibit their regulatory mechanism in an interaction or proximity dependent manner.

Q1:GLA(gene X, gene Y|enhancer Z)			
GO ID	***P*-value**	**GO term**	**Genes**
GO:0005923	< 0.0001	tight junction	CLDN14,IGSF5,JAM2,CXADR
GO:0030424	0.0033	axon	KCNJ6,NCAM2,DSCAM
GO:0030054	0.0013	cell junction	CLDN14,GRIK1,IGSF5,JAM2,ITSN1
GO:0006952	0.0007	defense response	MX2,ICOSLG,MX1
GO:0048663	0.0033	neuron fate commitment	OLIG1,OLIG2
GO:0004843	0.0007	ubiquitin-specific protease activity	USP16,USP25
Q2:GLA(enhancer X, enhancer Y|gene Z)			
GO ID	***P*-value**	**GO term**	**Genes**
GO:0007596	< 0.0001	blood coagulation	IFNAR2,IL10RB,IFNGR2,IFNAR1
GO:0005829	< 0.0001	cytosol	PFKL,PDE9A,BACH1,ITSN1,MX1,RRP1B,PCBP3, AGPAT3,DNMT3L,RSPH1,MCM3AP,PDXK,CRYZL1, PCP4,SIK1,CBS,PCNT,SOD1,C2CD2,UBASH3A, CBR3,UBE2G2
GO:0042776	< 0.0001	mitochondrial ATP synthesis coupled proton transport	ATP5O,ATP5J
GO:0005753	< 0.0001	mitochondrial proton-transporting ATP synthase complex	ATP5O,ATP5J
GO:0030688	< 0.0001	preribosome, small subunit precursor	RRP1B,RRP1
GO:0050733	< 0.0001	RS domain binding	SON,U2AF1
GO:0022857	< 0.0001	transmembrane transporter activity	ATP5O,ATP5J
GO:0004905	< 0.0001	type I interferon receptor activity	IFNAR2,IFNAR1
GO:0004090	0.0001	carbonyl reductase (NADPH) activity	CBR3,CBR1
GO:0006952	0.0001	defense response	ICOSLG,MX1,TFF3,MX2
GO:0009615	0.0001	response to virus	IFNAR2,IFNGR2,IFNAR1,MX1,MX2

Table II. : Enriched GO terms in Q1 and Q2.

2.4 The genes in liquid association with epigenetically modified enhancers have functions related to autoimmune diseases

To gain a better understanding of the functional significance of the genes in liquid association with enhancers with epigenetic marks, we performed gene ontology (GO) analysis using a previously published method that takes gene length and GC content into account when sampling for the control genes, and we obtain six and 10 enriched GO terms for Q1 and Q2 respectively (Table 2).

Among all the significant GO terms, defense response (GO:0006952) is enriched in both Q1 and Q2 analyses, and it has been previously identified as one of the enriched biological processes affected during differentiation of monocytes into macrophages through changes in the epigenetic and transcription state [Saeed et al. 2014]. In Javierre *et al.*, defense response is also an enriched GO term among genes that were identified to play a role in autoimmune diseases, where environmentally driven epigenetic changes are thought to contribute to the etiology [Javierre 2010].

2.5 Genes in liquid association with conserved enhancers are also more conserved

Previous studies showed that H3K4me1, H3K4me4, and H3K27ac have higher conservation levels between species [Woo and Li 2012]. This finding prompted us to evaluate the conservation of the corresponding associated genes of these three histone marks. We calculated the sequence conservation of the exonic regions of these genes using PhastCons, and the result shows that the genes in liquid association of the three conserved histone marks also demonstrate significantly higher conservation level than the background controls (Figure 6). This finding suggests that these three histone marks could facilitate the conserved epigenetic regulation of the targeted gene expression through enhancers.

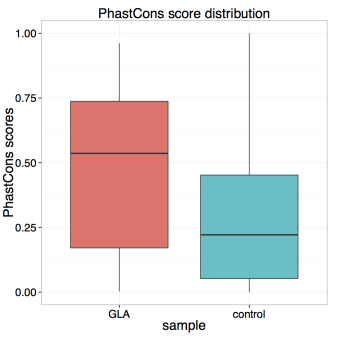


Fig. 6: PhastCons score distribution comparing genes from significant triplets involving conserved histone marks vs. the rest of the genes. GLA: 205 genes significantly associated with H3K4me1, H3K4me3, or H3K4ac. control: other genes in human chromosome 21. The PhastCons score of GLA is significantly higher than that of control (K-S test, P-value = 4.363E-07).

3. CONCLUSIONS

We were able to generate thousands of significant GLA triplets from the imputed epigenetic dataset. To find significance in these findings, we first examined the possible relationship between physical proximity and a liquid association relationship. Through 1-D distance and Hi-C data, we were able that significant liquid association triplets are closer, and interact more in certain histone mark enrichments. Notably, H3K27me3 triplets were closer, and interacted more in both Q1 and Q2. To further look at function of genes that have liquid association relationships, we performed GO analysis which showed the defense response term is enriched. This was consistent with previous work that epigenetic changes play a role in autoimmune diseases. Finally, we found that conservation of liquid association related genes was significant compared to random genes. Through these analyses, we show that liquid association has significant implications in terms of gene and enhancer function, conservation, and physical proximity.

4. FUTURE WORK

(1) Expand the gene, enhancer, and histone mark lists to the entire human genome for GLA calculations. When estimating the P-values, we will increase the permutation test to at least 1000 to obtain more accurate P-values. In addition, we will also perform false discovery rate (FDR) correction on the P-values to determine significant triplets.

(2) Include more Hi-C data from different cell types to investigate the interaction dependency on cell type-specificity between the significant GLA triplets.

(3) Determine the associated GWAS diseases of our genes in the significant triplets.

(4) Investigate further the functional and evolutionary implications of the genes dynamically associated with conserved histone marks.

5. METHODS

5.1 Selection of the triplets

We did two types of analyses, X and Y are genes vs. Z is an enhancer with epigenetic mark (Q1), and X and Y are enhancers each with epigenetic mark vs. Z is a gene (Q2). The enhancer coordinate information was obtained from the VISTA Enhancer Browser [Visel et al. 2007], and we overlapped the enhancer regions with all the given epigenetic marks to get the information of epigenetically modified enhancers. A mark intensity for one enhancer was calculated to be the summation of peak values overlapping the enhancer region normalized by the portion overlapped. The same method was also used for calculating the gene expression of a gene using the RNA-Seq data. In Q1 analysis, to reduce the number of comparisons, we first used the GeneFriends database to extract a list of gene with the gene expression poorly correlated with

Pearson's correlation coefficient between -0.1 and 0.1 and P-value of 0.01 in chromosome 21 as X and Y. For Q2, we used all possible combinations of the triplets.

5.2 GLA calculation and P-values estimation

We followed the GLA analysis procedures with default setting outlined in the GLA R package developed by Ho *et al.*, 2009 [Ho 2009]. Specifically, for expression GLA($X, Y|Z$), we calculated the GLA by separating the Z values into three ranges as suggested in the manual. The corresponding P-values for the triplets were estimated based on 100 permutation tests.

5.3 Distance between genes and associated enhancers

Distance between the genes and associated enhancers was calculated using the middle of the enhancer, and the gene start. The distance between the X and Z, and Y and Z were summed together as the distance metric. The distance score was calculated for all significant triplets, and all random triplets. Additionally, a mark specific score was calculated by only using the significant and random triplets that had one mark type.

5.4 Hi-C interaction between genes and enhancers

Hi-C data was obtained from the website of Ren Lab at University of California, San Diego (*http* : *//chromosome.sdsc.edu/mo*
c/download.html). We used their 40kb resolution Human ES Cell and Human IMR90 Hi-C data, which is normalized through a method developed by Yaffe *et al.* in 2011 [Yaffe 2011]. The interaction metric was calculated by summing the X-Z and Y-Z pair interaction score. The X-Z or Y-Z pair interaction score was found by taking the normalized interaction score in the 40kb bin that X and Z fell into. The middle of the enhancer or the gene start was used as the location of X, Y, or Z.

5.5 Gene ontology (GO) analysis

The GO terms of each gene were obtained from the Ensembl database. We used a previous published approach [Li et al. 2012] to identify enriched GO terms in our gene lists. Briefly, for each target gene in our list, we resampled a control from all the genes whose transcript length and GC content differences are within 5% of those of the target. We then calculated the frequency of each GO term in the control set. The number of control genes (denoted as X) for each GO term was determined, which was compared to that of the genes with significant GLA scores (denoted as x). This process was repeated 10,000 times to generate an expected frequency distribution for each GO term. An empirical P-value was calculated as $P = Pr(X \geq x)$. The significance cutoff for choosing enriched GO terms was 1/total GO terms considered.

5.6 PhastCons score calculation

To evaluate the conservation level of genes in Liquid Association with the previously shown conserved histone marks, H3K4me1, H3K4me3, and H3K27ac, we focused on triplets that have significant GLA scores with P-value < 0.01 ("significant triplets"). In Q1 analysis (X and Y are genes, and Z is enhancer), we identified three separate lists of genes from significant triplets associated with each of the three marks. In Q2 analysis (X and Y are enhancers, and Z is gene), we collected all significant triplets with histone mark of interest as X and/or Y and took the corresponding genes. For each Q1 and Q2, we took the union of all three gene lists as our testing genes. The rest of genes in chrosome 21 were the control genes. For both conserved histone mark-associated genes and the control genes in chromosome 21, we calculated the PhastCons scores extracted from the UCSC database (46-way PhastCons) [Siepel et al. 2005] for the exonic regions of each gene.

REFERENCES

Jason Ernst. 2015. Large-scale imputation of epigenomic datasets for systematic annotation of diverse human tissues. *Nat Biotech* (2015).

Michael Fernandez. 2012. Genome-Wide Enhancer Prediction from Epigenetic Signatures Using Genetic Algorithm-Optimized Support Vector Machines. *Nucleic Acids Research* (2012).

Nathaniel D Heintzman. 2009. Histone Modifications at Human Enhancers Reflect Global Cell Type-Specific Gene Expression. *Nature* (2009).

Yen-Yi Ho. 2009. GENERALIZED LIQUID ASSOCIATION. *Johns Hopkins University, Dept. of Biostatistics Working Papers* (2009).

Biola M Javierre. 2010. Changes in the Pattern of DNA Methylation Associate with Twin Discordance in Systemic Lupus Erythematosus. *Genome Research* (2010).

Gang Li, Jae Hoon Bahn, Jae-Hyung Lee, Guangdun Peng, Zugen Chen, Stanley F. Nelson, and Xinshu Xiao. 2012. Identification of allele-specific alternative mRNA processing via transcriptome sequencing. *Nucleic Acids Research* 40, 13 (2012), e104. DOI:http://dx.doi.org/10.1093/nar/gks280

Ker-Chau Li. 2002. Genome-wide coexpression dynamics: Theory and application. *Proceedings of the National Academy of Sciences* 99, 26 (2002), 16875–16880. DOI:http://dx.doi.org/10.1073/pnas.252466999

Sadia Saeed, Jessica Quintin, Hindrik H. D. Kerstens, Nagesha A. Rao, Ali Aghajanirefah, Filomena Matarese, Shih-Chin Cheng, Jacqueline Ratter, Kim Berentsen, Martijn A. van der Ent, Nilofar Sharifi, Eva M. Janssen-Megens, Menno Ter Huurne, Amit Mandoli, Tom van Schaik, Aylwin Ng, Frances Burden, Kate Downes, Mattia Frontini, Vinod Kumar, Evangelos J. Giamarellos-Bourboulis, Willem H. Ouwehand, Jos W. M. van der Meer, Leo A. B. Joosten, Cisca Wijmenga, Joost H. A. Martens, Ramnik J. Xavier, Colin Logie, Mihai G. Netea, and Hendrik G. Stunnenberg. 2014. Epigenetic programming of monocyte-to-macrophage differentiation and trained innate immunity. *Science* 345, 6204 (2014). DOI:http://dx.doi.org/10.1126/science.1251086

Daria Shlyueva. 2014. Transcriptional enhancers: from properties to genome-wide predictions. •*Nat Rev Genet* (2014).

Adam Siepel, Gill Bejerano, Jakob S. Pedersen, Angie S. Hinrichs, Minmei Hou, Kate Rosenbloom, Hiram Clawson, John Spieth, LaDeana W. Hillier, Stephen Richards, George M. Weinstock, Richard K. Wilson, Richard A. Gibbs, W. James Kent, Webb Miller, and David Haussler. 2005. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Research* 15, 8 (2005), 1034–1050. DOI:http://dx.doi.org/10.1101/gr.3715005

Sipko van Dam, Thomas Craig, and João Pedro de Magalhães. 2014. GeneFriends: a human RNA-seq-based gene and transcript co-expression database. *Nucleic Acids Research* (2014). DOI:http://dx.doi.org/10.1093/nar/gku1042

Axel Visel, Simon Minovitsky, Inna Dubchak, and Len A. Pennacchio. 2007. VISTA Enhancer Browser—a database of tissue-specific human enhancers. *Nucleic Acids Research* 35, suppl 1 (2007), D88–D92. DOI:http://dx.doi.org/10.1093/nar/gkl822

Yong H. Woo and Wen-Hsiung Li. 2012. Evolutionary Conservation of Histone Modifications in Mammals. *Molecular Biology and Evolution* 29, 7 (2012), 1757–1767. DOI:http://dx.doi.org/10.1093/molbev/mss022

Eitan Yaffe. 2011. Probabilistic modeling of Hi-C contact maps eliminates systematic biases to characterize global chromosomal architecture. *Nat Genet* (2011).

Received June 2015; revised June 2015; accepted June 2015