G&I Genomics & Informatics

eISSN 2234-0742
Genomics Inform 2015;13(2):40-44
http://dx.doi.org/10.5808/GI.2015.13.2.40

ORIGINAL ARTICLE

Genetic Architecture of Transcription and Chromatin Regulation

Kwoneel Kim, Hyoeun Bang, Kibaick Lee, Jung Kyoon Choi*

Department of Bio and Brain Engineering, Korea Advanced Institute of Science and Technology (KAIST), Daejeon 305-701, Korea

DNA microarray and next-generation sequencing provide data that can be used for the genetic analysis of multiple quantitative traits such as gene expression levels, transcription factor binding profiles, and epigenetic signatures. In particular, chromatin opening is tightly coupled with gene transcription. To understand how these two processes are genetically regulated and associated with each other, we examined the changes of chromatin accessibility and gene expression in response to genetic variation by means of quantitative trait loci mapping. Regulatory patterns commonly observed in yeast and human across different technical platforms and experimental designs suggest a higher genetic complexity of transcription regulation in contrast to a more robust genetic architecture of chromatin regulation.

Keywords: epigenetic process, gene expression, quantitative trait loci, regulatory regions

Introduction

Quantitative trait loci (QTL) mapping has been widely used to discover underlying genetic factors that can explain particular phenotypes of interest. Thanks to DNA microarray technology, the expression phenotype of thousands of genes was associated with the genotypes across the whole genome in expression QTL mapping [1-7]. Recent advent of next-generation sequencing technology has enabled a genetic profiling of chromatin traits as well as more in-depth analyses of gene expression variation. Especially, the mechanisms controlling chromatin accessibility have been of particular interest because of their importance in a wide spectrum of DNA regulation processes. For example, Degner *et al*. [8] utilized DNaseI hypersensitivity assay coupled with high-throughput sequencing (DNase-seq) to map genome-wide chromatin accessibility to the genotypes across 70 human individuals, for which a previous RNA-sequencing (RNA-seq)-based expression QTL study revealed new insights into the genetic regulation of transcription [5]. In yeast, expression QTL mapping has been performed for the genetic dissection of transcription regulation in a cross between two parental strains of *Saccharomyces cerevisiae* (BY4716 and RM11-1a) based on DNA microarrays [1, 2, 9-11]. In order to generate a matched dataset of chromatin accessibility for this set of yeast individuals, we carried out Formaldehyde-Assisted Isolation of Regulatory Elements followed by sequencing (FAIRE-seq) for a total of 96 segregants from the cross of BY4716 and RM11-1a [12].

In this study, we sought to dissect the genetic architecture of the regulation of gene expression and chromatin accessibility by analysing previous data generated in yeast and human based on different technical platforms and experimental designs. Our main goal was to find differences in the overall regulatory structure between open chromatin and gene expression. We were also interested to determine whether the two distant species, namely yeast and human, would be different in genetic regulatory architecture and to estimate the effect of the technical or experimental differences in genotyping and measuring the quantitative traits.

Methods

Processing of human genotype data

Genotype data from the HapMap project [13] and 1000 Genomes Project [14] for 70 Yoruba (YRI) lymphoblastoid

Received May 7, 2015; Revised June 10, 2015; Accepted June 11, 2015

***Corresponding author:** Tel: +82-42-350-4327, Fax: +82-42-350-4310, E-mail: jungkyoon@kaist.ac.kr

Copyright © 2015 by the Korea Genome Organization
It is identical to the Creative Commons Attribution Non-Commercial License (http://creativecommons.org/licenses/by-nc/4.0/).

cell lines were used for DNase-seq analysis [8]. The genotype of each single nucleotide polymorphism (SNP) locus was estimated based on the Bayesian framework by means of the BIMBAM tool [15] and the genotype estimates were made available at http://eqtl.uchicago.edu/dsQTL_data/GENOTYPES/. We first selected 2,157,286 genetic markers (SNPs) with the minor allele frequency greater than 30%. To reduce complexity and ease interpretation, we focused on the genetic variants that can change the function of the protein (non-synonymous SNPs) or the abundance of the protein (SNPs associated with the expression level of a nearby gene). The SIFT tool [16] was used to identify non-synonymous SNPs. We performed expression QTL mapping as described below and identified SNPs that were associated ($p < 10^{-5}$) in *cis* (within 200 kb from the nearest gene). Taken together, 7,211 SNPs were identified for QTL mapping.

Processing of human gene expression data

RNA-seq data for 69 YRI lymphoblastoid cell lines [5] were downloaded from http://eqtl.uchicago.edu/RNA_Seq_data/results. A total of 18,147 genes were used after normalization to zero mean and unit variance.

Processing of human chromatin accessibility data

DNase-seq data for 70 YRI lymphoblastoid cell lines [8] were downloaded from http://eqtl.uchicago.edu/dsQTL_data/MAPPED_READS/. Sequence reads from multiple replicates for each sample were combined and F-Seq [17] was run to identify the peaks of the reads from each sample. Statistical significance of the peak was determined by fitting the data to a gamma distribution to obtain the p-value (script obtained from the F-Seq authors). $p < 10^{-3}$ was used to identify significant peaks from each sample. The overlapping peaks across the YRI individuals were merged into a single peak by using the mergeBED command of BEDTools [18], resulting in a total of 265,130 accessible chromatin regions. For each sample, the number of the DNase-seq reads mapped to each region was counted and the read count was normalized as previously suggested [19, 20] to obtain normalized chromatin accessibility, which was then further normalized to zero mean and unit variance across the YRI samples. Accessible regions falling on promoters or enhancers were identified based on chromatin annotation by Ernst *et al*. [21]. A total of 45,781 chromatin regions were found to reside in active promoters, weak promoters, poised promoters, strong enhancers, and weak enhancers annotated in the GM12878 lymphoblastoid cell line.

Processing of yeast data

Genotype and gene expression microarray data [10] used in previous expression QTL studies [1, 2, 9] for >100 segregants from a cross between two parental strains of yeast (BY4716 and RM11-1a) were obtained. As previously suggested [22], adjacent genetic markers with less than three genotypic mismatches across the yeast strains were merged into the average genotype profile, resulting in 1,533 unique markers. We employed the microarray dataset of normalized expression levels of 5,352 genes as previously used [10]. FAIRE experiments were performed based on the published protocol [23]. The FAIRE-seq data for the 96 yeast strains from our previous work [12] is available at the Gene Expression Omnibus (GEO) database with accession number GSE33466. Briefly, we identified open chromatin regions in 96 yeast segregants by means of F-Seq [17]. The overlapping peaks across the 96 strains were merged into a single peak by using BEDTools [18], resulting in a total of 7,527 accessible chromatin regions. For each sample, the number of the FAIRE-seq reads mapped to each region was counted and the read count was normalized as previously suggested [19, 20] to obtain normalized chromatin accessibility, which was then further normalized to zero mean and unit variance across the 96 segregants.

Trans-QTL analysis

Sixty-three human samples were commonly present in the RNA-seq [5] and DNase-seq [8] data, and 96 yeast segregants in the microarray and FAIRE-seq data. Therefore, we used the common samples for our QTL mapping. Linear regression was carried out leading to 2,100,341 chromatin associations and 975,333 expression associations in human,

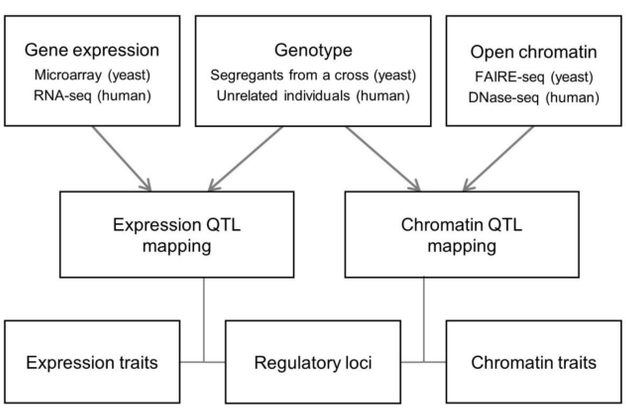

Fig. 1. Data analysis scheme. Relationships between the genetic regulatory loci and the quantitative traits (chromatin accessibility or gene expression) were explored in yeast and human. Data from different experimental settings and technical platforms were integrated into a unified analytical framework. RNA-seq, RNA-sequencing; FAIRE-seq, Formaldehyde-Assisted Isolation of Regulatory Elements followed by sequencing; DNase-seq, DNaseI hypersensitivity assay coupled with high-throughput sequencing; QTL, quantitative trait loci.

and 110,802 chromatin linkages and 164,217 expression linkages in yeast at $p < 0.01$. Genetic markers farther than 200 kb from the nearest gene in human and 100 kb in yeast were identified to examine regulatory relationships acting in *trans*. For each trait (gene or accessible chromatin region) and regulatory locus (genetic marker), the number of associations or linkages was obtained and divided by the total number of associations at a given p-value.

Results and Discussion

For a systematic comparative analysis of chromatin regulation and transcription regulation, we carried out QTL mapping in a unified analytical framework by applying the

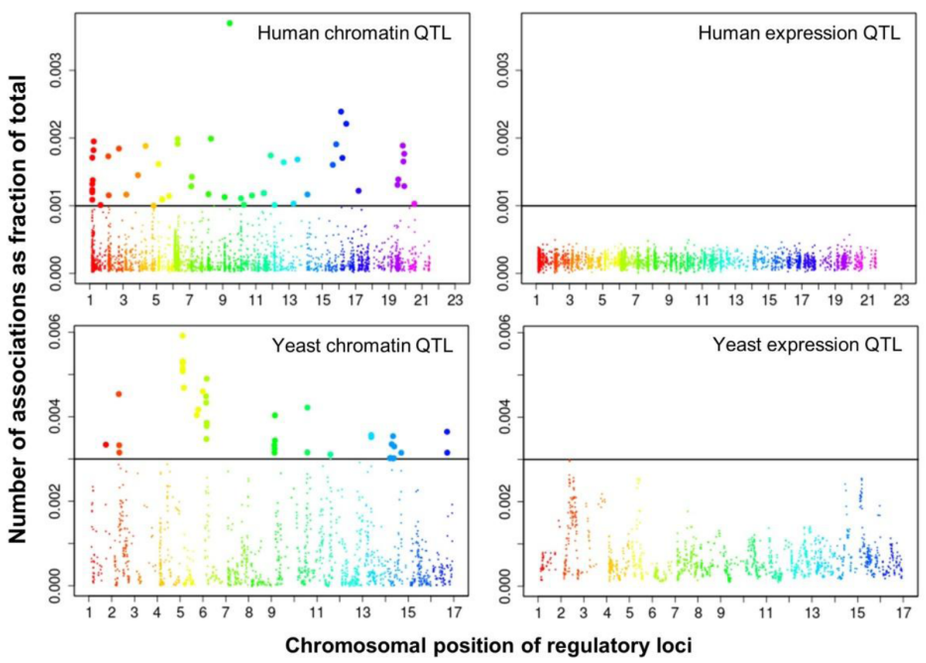

Fig. 2. Manhattan plots displaying the chromosomal distribution of regulatory loci and (x axis) and the number of associations for each locus divided by the total number of associations (i.e., percentage on the y axis) at a given p-value in human ($p < 0.01$) and yeast ($p < 0.05$). QTL, quantitative trait loci.

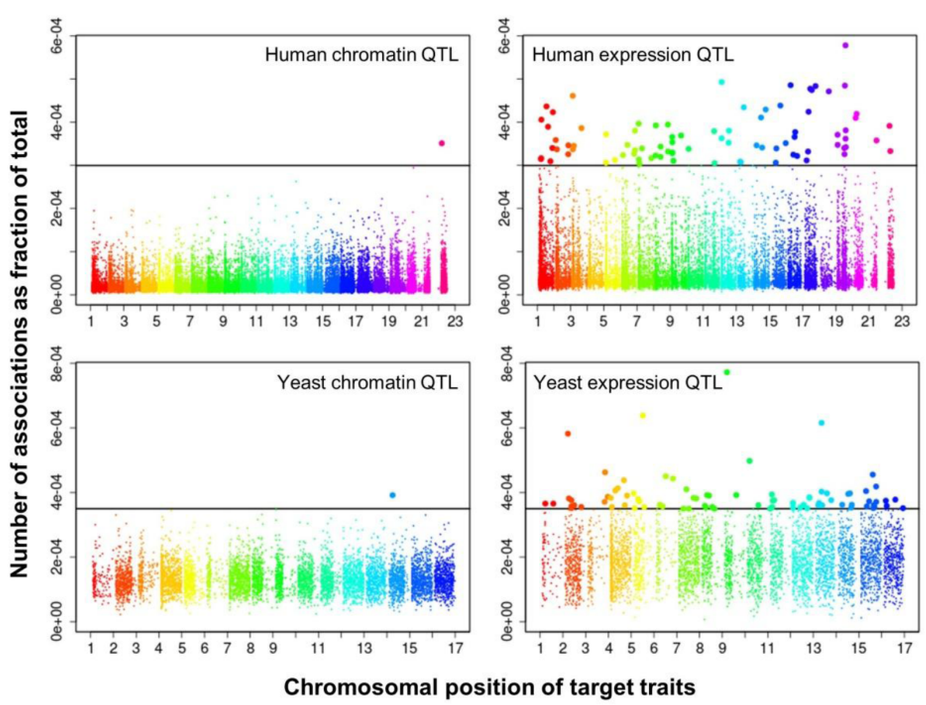

Fig. 3. Manhattan plots displaying the chromosomal distribution of target traits (x axis) and the number of associations for each trait divided by the total number of associations (i.e., percentage on the y axis) at a given p-value in human ($p < 0.01$) and yeast ($p < 0.05$). QTL, quantitative trait loci.

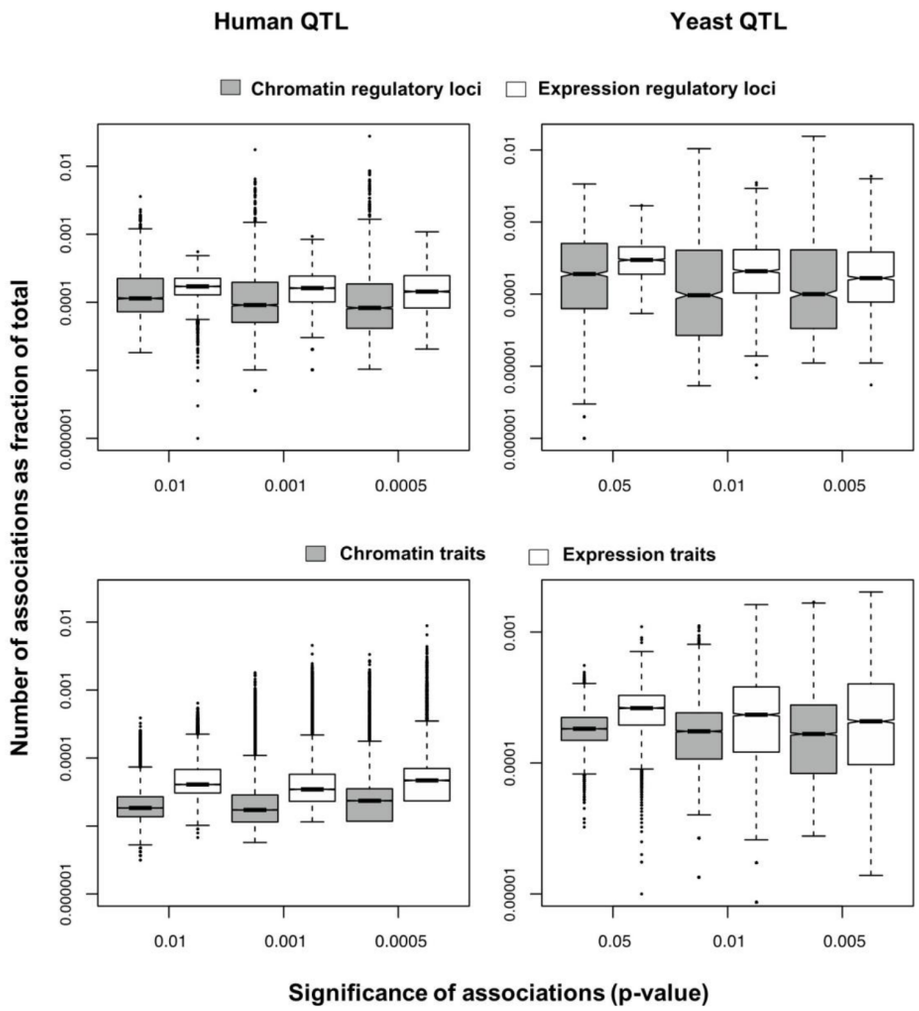

Fig. 4. Box plots showing the percentage of associations for each regulatory locus (upper) or quantitative trait (lower) from chromatin quantitative trait loci (QTL) and expression QTL mapping on a log scale in human (left) and yeast (right).

identical statistical methods and criteria for open chromatin and gene expression data in yeast and human (Methods and Fig. 1). From the human chromatin accessibility data [8], we selected chromatin sites at the promoter or enhancer (as defined by Ernst *et al*. [21]). Yeast and human QTL mapping was different in the genetics setting (linkage vs. association) and technical platforms (DNA microarray vs. RNA-seq and FAIRE-seq vs. DNase-seq) (Fig. 1).

We selected significant *trans*-associations based on the p-value of linear regression and then examined the distribution of linkages between regulatory loci (genetic markers) and quantitative traits (gene expression levels or chromatin accessibility). First, there were particular regulatory loci that were associated with a large number of target chromatin traits but not with gene expression traits both in human and yeast (Fig. 2). Second, expression traits than chromatin traits were associated with a greater number of regulatory loci (Fig. 3). This trend was consistently found when varying p-values were used (Fig. 4). Intriguingly, the number of targets for chromatin regulatory loci is more dispersed resulting in outliers with extremely many targets while the average number is lower as compared to the pattern of expression regulatory loci (grey vs. white in the upper panel of Fig. 4). On the other hand, the number of associated regulatory loci was much greater for gene expression traits in terms of the outliers and average as well (grey vs. white in the lower panel of Fig. 4).

These findings can be interpreted as follows. There are a few regulatory hotspots that are responsible for a large number of chromatin sites. Most regulatory loci, however, tend to influence fewer chromatin traits than gene expression traits. Different biological functions are expected between the promiscuous and specific chromatin regulators. On the other hand, gene expression seems to be more reactive to genetic variation when judged by the number of associated regulatory loci. This may imply that transcription processes receive more regulatory inputs than chromatin regulation, in which case chromatin accessibility may not be

a good predictor of precise gene expression levels. There was a remarkable consistency of the data for two evolutionarily distant species across different technical platforms and experimental settings, thereby supporting the biological implications of the findings while ruling out the possibility of technical artifacts.

Based on these findings, we propose that chromatin regulation mechanisms have evolved a stable genetic architecture while the transcription regulatory network maintains high genetic complexity and connectivity and is more susceptible to mutations. First, the lower number of associations per chromatin trait suggests that chromatin structure is more robust to genetic perturbation. On the contrary, gene expression traits have more associations, increasing the probability of targets being associated with mutations. Second, chromatin associations are biased to a handful of loci responsible for an extremely large number of chromatin sites. As long as these hotspot loci are protected from mutations, the perturbation of chromatin regulation architecture could be minimized. In contrast, the average number of associations is high for transcription regulatory loci, increasing the overall effect of single mutations. Given the higher probability of expression traits being associated with mutations and the greater influence of individual mutations on transcription, it is conceivable that phenotypic diversity is more likely to be created in response to genetic variation at the level of transcription processes downstream of chromatin regulation steps.

Acknowledgments

This work was supported by the KAIST Future Systems Healthcare Project and by a grant [2013M3A9C4078139] from the National Research Foundation of Korea. This work was also supported by Brain Pool Program through the Korean Federation of Science and Technology Societies (141S-4-3-0035).

References

1. Brem RB, Yvert G, Clinton R, Kruglyak L. Genetic dissection of transcriptional regulation in budding yeast. *Science* 2002; 296:752-755.
2. Brem RB, Storey JD, Whittle J, Kruglyak L. Genetic interactions between polymorphisms that affect gene expression in yeast. *Nature* 2005;436:701-703.
3. Cheung VG, Spielman RS, Ewens KG, Weber TM, Morley M, Burdick JT. Mapping determinants of human gene expression by regional and genome-wide association. *Nature* 2005;437: 1365-1369.
4. Morley M, Molony CM, Weber TM, Devlin JL, Ewens KG, Spielman RS, *et al*. Genetic analysis of genome-wide variation in human gene expression. *Nature* 2004;430:743-747.
5. Pickrell JK, Marioni JC, Pai AA, Degner JF, Engelhardt BE, Nkadori E, *et al*. Understanding mechanisms underlying human gene expression variation with RNA sequencing. *Nature* 2010;464:768-772.
6. Rockman MV, Kruglyak L. Genetics of global gene expression. *Nat Rev Genet* 2006;7:862-872.
7. Schadt EE, Monks SA, Drake TA, Lusis AJ, Che N, Colinayo V, *et al*. Genetics of gene expression surveyed in maize, mouse and man. *Nature* 2003;422:297-302.
8. Degner JF, Pai AA, Pique-Regi R, Veyrieras JB, Gaffney DJ, Pickrell JK, *et al*. DNase I sensitivity QTLs are a major determinant of human expression variation. *Nature* 2012;482:390-394.
9. Yvert G, Brem RB, Whittle J, Akey JM, Foss E, Smith EN, *et al*. Trans-acting regulatory variation in Saccharomyces cerevisiae and the role of transcription factors. *Nat Genet* 2003;35:57-64.
10. Brem RB, Kruglyak L. The landscape of genetic complexity across 5,700 gene expression traits in yeast. *Proc Natl Acad Sci U S A* 2005;102:1572-1577.
11. Choi JK, Kim YJ. Epigenetic regulation and the variability of gene expression. *Nat Genet* 2008;40:141-147.
12. Lee K, Kim SC, Jung I, Kim K, Seo J, Lee HS, *et al*. Genetic landscape of open chromatin in yeast. *PLoS Genet* 2013;9: e1003229.
13. International HapMap Consortium, Frazer KA, Ballinger DG, Cox DR, Hinds DA, Stuve LL, *et al*. A second generation human haplotype map of over 3.1 million SNPs. *Nature* 2007;449:851-861.
14. 1000 Genomes Project Consortium, Abecasis GR, Altshuler D, Auton A, Brooks LD, Durbin RM, *et al*. A map of human genome variation from population-scale sequencing. *Nature* 2010;467:1061-1073.
15. Guan Y, Stephens M. Practical issues in imputation-based association mapping. *PLoS Genet* 2008;4:e1000279.
16. Ng PC, Henikoff S. SIFT: Predicting amino acid changes that affect protein function. *Nucleic Acids Res* 2003;31:3812-3814.
17. Boyle AP, Guinney J, Crawford GE, Furey TS. F-Seq: a feature density estimator for high-throughput sequence tags. *Bioinformatics* 2008;24:2537-2538.
18. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* 2010;26:841-842.
19. Choi JK. Contrasting chromatin organization of CpG islands and exons in the human genome. *Genome Biol* 2010;11:R70.
20. Choi JK, Bae JB, Lyu J, Kim TY, Kim YJ. Nucleosome deposition and DNA methylation at coding region boundaries. *Genome Biol* 2009;10:R89.
21. Ernst J, Kheradpour P, Mikkelsen TS, Shoresh N, Ward LD, Epstein CB, *et al*. Mapping and analysis of chromatin state dynamics in nine human cell types. *Nature* 2011;473:43-49.
22. Lee SI, Pe'er D, Dudley AM, Church GM, Koller D. Identifying regulatory mechanisms using individual variation reveals key role for chromatin modification. *Proc Natl Acad Sci U S A* 2006;103:14062-14067.
23. Giresi PG, Kim J, McDaniell RM, Iyer VR, Lieb JD. FAIRE (Formaldehyde-Assisted Isolation of Regulatory Elements) isolates active regulatory elements from human chromatin. *Genome Res* 2007;17:877-885.