**Our experience**

In sum, this project will result in a large and comprehensive dataset of comparable EGIs from many species, as well as functional data such as scRNA-seq data, of multiple cell types and species. Analyses of such datasets will substantially improve our understanding of the evolution of enhancers and EGIs. In turn, these results will elucidate how enhancers regulate gene expression and paralog generation. Taken together, these discoveries will be enormously useful for studying human health and disease. Particularly, EGIs that are conserved across mammals could represent potential targets for disease-causing mutations. Such conserved EGIs cannot be identified from sequence comparisons or non-comparative analyses. We have extensive experience with processing and analyzing Hi-C data(*1, 2*), identifying enhancers(*3*) and their target genes(*4, 5*), and conducting comparative analyses(*6, 7*), for which we have developed many pipelines that are readily available for this project(*6, 8*).

From whole-genome sequencing of multiple dogs and wolves, Axelsson et al. analyzed the resultant single nucleotide variants (SNV) and found that ten genes with key roles in starch digestion and fat metabolism show signals of selection in dogs(*9*). We analyzed SNVs between dog and wolf, and found that these SNVs are enriched in liver-specific enhancers. The genes close to the enhancers with SNV enrichment tend to work in biological pathways such as “cellular response to toxic substance” (Bonferroni p-value = 4.8E-6, FDR=0.038), “positive regulation of response to external stimulus” (Bonferroni p-value = 5.5E-6, FDR=0.038), and “wound healing” (Bonferroni p-value = 5.4E6, FDR=0.038). Accurate EGIs will provide better SNV-gene linkage for such functional analyses. Moreover, compared to individual SNVs, EGIs are a more powerful and direct regulator of gene expression. Therefore, we will use the systematically identified EGIs to probe for their connections with gene constraints and functions.

**Background Introduction**

**Regulatory function and health importance of enhancers**

An enhancer is a stretch of DNA (~100-1,000 bp) that dramatically boosts the transcription of its target genes(*10*). Enhancers can regulate gene expression independent of their orientation, location, and distance to the gene(*11, 12*). Enhancers as cis-regulatory elements (CREs) were first discovered nearly 40 years ago in simian virus 40, and later found in almost all eukaryotic genomes(*11, 13*). A typical mammalian gene is controlled by several enhancers, which can fine-tune gene expression based on developmental stage, cell type, and signaling cues(*14*). As a result, enhancers play central roles in many phenotypes and human diseases. Consistent, enhancers tend to be enriched for disease-associated genetic variants(*15*). Roughly 847 disease-associated enhancers have been reported in 143 human diseases in the DiseaseEnhancer database(*16*), and more are expected to be discovered given that the field is developing rapidly. Therefore, understanding how enhancers modulate gene expression and identifying conserved EGIs in humans will be enormously important to human health.

**Identifying enhancers and enhancer-gene interactions (EGIs)** **at a genomic scale**

Many efforts have been made over the past several decades to identify enhancers. Initially, the highly conserved non-coding sequences in the genome were identified as enhancers(*17-19*). About half of the hundreds of conserved sequences identified can activate gene expression in transgenic mouse reporter assays(*18, 20*). However, a substantial number of enhancers have modest or no conservation across species(*21, 22*); such enhancers would be missed by conservation detection. Moreover, conservation alone does not indicate at which developmental stage and cell type the enhancer carries out its function. To address these gaps, another approach to identify enhancers is by the epigenetic marks associated with enhancer functions, such as p300 binding(*23*), H3K4me1(*24*), and H3K27ac(*25*). These and other histone marks have proven useful for enhancer identification by *in vitro*(*24*) or *in vivo*(*23*) experimental validation. The epigenetic marks have been integrated by machine learning, especially deep learning, algorithms to further improve enhancer detection(*26*). However, these many histone marks are mainly determined for human and mouse cell types.

Compared to enhancer detection, it is even more difficult to identify EGIs because they are mediated by genome looping, and thus require knowing the contacts of genomics regions in three-dimensional (3D) space. Such contacts at a genomic scale were first identified using the Hi-C technique(*27*); however, the 5-20 kb resolution of this method is usually much longer than an enhancer or sometimes even a gene, and thus usually cannot identify individual EGIs(*28*). To this end, capture Hi-C (CHi-C) combines Hi-C with capture probes usually targeting promoters to focus only on the genomic contacts with promoters involved, and thus increase the average resolution to ~3 kb on average using a six-cutter enzyme(*29, 30*). Probes targeting enhancers can be also designed according to enhancer sequences to capture enhancer-enhancer interactions. With a four-cutter enzyme, we successfully identified EGIs in mouse(*31*) and human(*32*) at an average ~1 kb resolution. A DNase Hi-C protocol have been developed to obtain even higher enhancer resolution (less than 1 kb)(*33*). Moreover, because this new protocol uses DNase instead of RE to fragmentize genomes, this avoids potential differences in enzyme site frequencies across organisms, rendering the EGIs more comparable.

**Evolution of enhancers and EGIs**

Researchers have been intrigued by the evolution of enhancers and EGIs since they were first discovered. Case studies have demonstrated that by gain or loss of enhancers, a gene changes its expression level and developmental time to evolve new phenotypes(*34*). However, the evolutionary rate of EGIs at a genomic scale is unknown due to a lack of comparable data across multiple species. Although many enhancers and EGIs have been identified by high-throughput methods such as chromatin immunoprecipitation sequencing (ChIP-seq), Hi-C(*27*), and CHi-C(*30, 31, 35*), most are restricted to human and mouse. Moreover, the known EGIs were generated from different tissues by different labs with different protocols. These mismatches and potential batch effects further impede comparative analyses. In previous work, Villar et al. used H3K27ac as the signal of active enhancers to identify enhancers in the liver of humans and other mammals(*36*). These researchers observed that in mammalian genomes enhancers evolve rapidly by ancestral DNA exaptation. To identify EGIs, due to a lack of 3D genomic data, they made a compromise and assigned enhancers to neighboring genes. A large fraction of the EGIs are expected to be false positives or negatives. From this dataset, Villar and colleagues found that recently evolved enhancers could be associated with genes under positive selection(*36*). However, this weak trend is likely due to potential biases in the EGI assignment.

**Impact of enhancers on paralog formation**

Gene duplications are a major source of new genes and ultimately of new biological functions(*37*). Upon duplication, the two copies are expected to increase the dosage of the gene product, and this dosage imbalance is usually deleterious. And thus one of the copies will be purged out quickly by selection. Even the duplicate is not deleterious and reach fixation in the population by chance. This duplicate is subject to neutral evolution and eventually becomes a pseudogene. Occasionally, the two copies become stable paralogs by neofunctionalization (one copy gains new functions)(*38*) or subfunctionalization (the two copies divide the ancestral functions between them)(*38-41*). Increasing evidence indicates that, in many cases, the two paralogs reduce their respective expression to restore the dosage balance(*39-41*), which is a special case of subfunctionalization. The formation of such paralogs can be explained by the duplication-degeneration-complementation process, in which the two copies acquire mutations in the CREs to reduce their respective expression and thus retain dosage balance(*39*).

Lan et al. proposed that the dosage balance could be maintained by enhancer sharing between the two copies(*41*). The divided usage of enhancers reduces the expression of both copies. As supportive evidence, the authors observed that – in the GM12878 cell line – the promoters of neighboring paralogs in the human genome have more contacts in terms of Hi-C reads than neighboring singleton genes(*41*). This suggests that paralog genes tend to be co-regulated by enhancers. However, the difference in contacts is quite small, which could be due to the low resolution of Hi-C data or other confounding factors. For example, with similar promoters, neighboring paralogs are more likely to express with the same spatial and temporal pattern than neighboring singletons, and thus paralog promoters are expected to have more contacts by chance. In addition, due to the low resolution of Hi-C data, the paralog pairs these authors analyzed are at least 20 kb away in the chromosome; thus, many informative pairs might be excluded from these analyses.

**Impact of enhancers on gene expression levels and dynamics**

It is clear that enhancers activate and increase the transcription of a gene. Many studies indicate that enhancers serve as centers for the assembly of the pre-initiation complex for transcription(*42-46*). Moreover, some case studies also suggest that enhancers have a role in RNA polymerase II elongation(*47-49*). Researchers have studied the impact of the existing EGIs on gene expression in particular cell types. However, it is still controversial how multiple enhancers collaboratively modulate gene expression(*50-52*). Subsequently, Osterwalder et al. found that deleting one or even several enhancers of a gene results in unnoticeable or subtle phenotypic changes in mice(*53*), suggesting enhancer redundancy. However, redundant enhancers are subject to neutral evolution and thus are expected to deteriorate rapidly.

**Generalizability and potential contribution of the project**

Enhancers play a pivotal role in gene expression. In addition, enhancers evolve much more rapidly than promoters. These together indicate that enhancers can contribute substantially to gene expression changes(*36*). However, to modulate gene expression, enhancers as distal regulatory elements need to interact with their target genes by DNA looping(*54*). Our CHi-C data can identify the set of enhancers used by each gene in the genome. An understanding of gene regulation by enhancers is important to human development and health, as differences in gene regulation are a primary cause responsible for human physiology and behavior(*55, 56*). Moreover, many enhancers and EGIs have been identified in human diseases(*16*).

**REFERENCES**

1. K. K. Yan, S. Lou, M. Gerstein, MrTADFinder: A network modularity based approach to identify topologically associating domains in multiple resolutions. *PLoS Comput Biol* **13**, e1005647 (2017).

2. K. K. Yan, G. G. Yardimci, C. Yan, W. S. Noble, M. Gerstein, HiC-spector: a matrix library for spectral and reproducibility analysis of Hi-C contact maps. *Bioinformatics* **33**, 2199-2201 (2017).

3. A. Sethi *et al.*, A cross-organism framework for supervised enhancer prediction with epigenetic pattern recognition and targeted validation. *bioRxiv*, 385237 (2018).

4. Y. Fu *et al.*, FunSeq2: a framework for prioritizing noncoding regulatory variants in cancer. *Genome Biol* **15**, 480 (2014).

5. Q. Cao *et al.*, Reconstruction of enhancer-target networks in 935 samples of human primary cells, tissues and cell lines. *Nat Genet* **49**, 1428-1436 (2017).

6. K. K. Yan *et al.*, OrthoClust: an orthology-based network framework for clustering data across multiple species. *Genome Biol* **15**, R100 (2014).

7. C. Sisu *et al.*, Comparative analysis of pseudogenes across three phyla. *Proc Natl Acad Sci U S A* **111**, 13361-13366 (2014).

8. A. Anil, R. Spalinskas, O. Akerborg, P. Sahlen, HiCapTools: a software suite for probe design and proximity detection for targeted chromosome conformation capture applications. *Bioinformatics* **34**, 675-677 (2018).

9. E. Axelsson *et al.*, The genomic signature of dog domestication reveals adaptation to a starch-rich diet. *Nature* **495**, 360-364 (2013).

10. H. K. Long, S. L. Prescott, J. Wysocka, Ever-Changing Landscapes: Transcriptional Enhancers in Development and Evolution. *Cell* **167**, 1170-1187 (2016).

11. J. Banerji, S. Rusconi, W. Schaffner, Expression of a beta-globin gene is enhanced by remote SV40 DNA sequences. *Cell* **27**, 299-308 (1981).

12. C. T. Ong, V. G. Corces, Enhancer function: new insights into the regulation of tissue-specific gene expression. *Nat Rev Genet* **12**, 283-293 (2011).

13. P. Moreau *et al.*, The SV40 72 base repair repeat has a striking effect on gene expression both in SV40 and other chimeric recombinants. *Nucleic Acids Res* **9**, 6047-6068 (1981).

14. W. Schaffner, Enhancers, enhancers - from their discovery to today's universe of transcription enhancers. *Biol Chem* **396**, 311-327 (2015).

15. K. K. Farh *et al.*, Genetic and epigenetic fine mapping of causal autoimmune disease variants. *Nature* **518**, 337-343 (2015).

16. G. Zhang *et al.*, DiseaseEnhancer: a resource of human disease-associated enhancer catalog. *Nucleic Acids Res* **46**, D78-D84 (2018).

17. G. Bejerano *et al.*, Ultraconserved elements in the human genome. *Science* **304**, 1321-1325 (2004).

18. L. A. Pennacchio *et al.*, In vivo enhancer analysis of human conserved non-coding sequences. *Nature* **444**, 499-502 (2006).

19. A. Visel, J. Bristow, L. A. Pennacchio, Enhancer identification through comparative genomics. *Semin Cell Dev Biol* **18**, 140-152 (2007).

20. A. Visel *et al.*, Ultraconservation identifies a small subset of extremely constrained developmental enhancers. *Nat Genet* **40**, 158-160 (2008).

21. D. Schmidt *et al.*, Five-vertebrate ChIP-seq reveals the evolutionary dynamics of transcription factor binding. *Science* **328**, 1036-1040 (2010).

22. M. J. Blow *et al.*, ChIP-Seq identification of weakly conserved heart enhancers. *Nat Genet* **42**, 806-810 (2010).

23. A. Visel *et al.*, ChIP-seq accurately predicts tissue-specific activity of enhancers. *Nature* **457**, 854-858 (2009).

24. N. D. Heintzman *et al.*, Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. *Nat Genet* **39**, 311-318 (2007).

25. M. P. Creyghton *et al.*, Histone H3K27ac separates active from poised enhancers and predicts developmental state. *Proc Natl Acad Sci U S A* **107**, 21931-21936 (2010).

26. Y. Li, W. Shi, W. W. Wasserman, Genome-wide prediction of cis-regulatory regions using supervised deep learning methods. *BMC Bioinformatics* **19**, 202 (2018).

27. E. Lieberman-Aiden *et al.*, Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science* **326**, 289-293 (2009).

28. F. Jin *et al.*, A high-resolution map of the three-dimensional chromatin interactome in human cells. *Nature* **503**, 290-294 (2013).

29. S. Schoenfelder *et al.*, The pluripotent regulatory circuitry connecting promoters to their long-range interacting elements. *Genome Res* **25**, 582-597 (2015).

30. B. Mifsud *et al.*, Mapping long-range promoter contacts in human cells with high-resolution capture Hi-C. *Nat Genet* **47**, 598-606 (2015).

31. P. Sahlen *et al.*, Genome-wide mapping of promoter-anchored interactions with close to single-enhancer resolution. *Genome Biol* **16**, 156 (2015).

32. Ö. Åkerborg *et al.*, High-resolution regulatory maps connect cardiovascular risk variants to disease related pathways. *bioRxiv*, 376699 (2018).

33. W. Ma *et al.*, Using DNase Hi-C techniques to map global and local three-dimensional genome architecture at high resolution. *Methods* **142**, 59-73 (2018).

34. J. D. Cande, V. S. Chopra, M. Levine, Evolving enhancer-promoter interactions within the tinman complex of the flour beetle, Tribolium castaneum. *Development* **136**, 3153-3160 (2009).

35. R. Jager *et al.*, Capture Hi-C identifies the chromatin interactome of colorectal cancer risk loci. *Nat Commun* **6**, 6178 (2015).

36. D. Villar *et al.*, Enhancer evolution across 20 mammalian species. *Cell* **160**, 554-566 (2015).

37. G. C. Conant, K. H. Wolfe, Turning a hobby into a job: how duplicated genes find new functions. *Nat Rev Genet* **9**, 938-950 (2008).

38. S. Ohno, *Evolution by Gene Duplication*. (Springer, Berlin, Heidelberg, 1970).

39. A. Force *et al.*, Preservation of duplicate genes by complementary, degenerative mutations. *Genetics* **151**, 1531-1545 (1999).

40. W. Qian, B. Y. Liao, A. Y. Chang, J. Zhang, Maintenance of duplicate genes and their functional redundancy by reduced expression. *Trends Genet* **26**, 425-430 (2010).

41. X. Lan, J. K. Pritchard, Coregulation of tandem duplicate genes slows evolution of subfunctionalization in mammals. *Science* **352**, 1009-1013 (2016).

42. F. Koch *et al.*, Transcription initiation platforms and GTF recruitment at tissue-specific enhancers and promoters. *Nat Struct Mol Biol* **18**, 956-963 (2011).

43. M. H. Kagey *et al.*, Mediator and cohesin connect gene expression and chromatin architecture. *Nature* **467**, 430-435 (2010).

44. M. Stumpf *et al.*, The mediator complex functions as a coactivator for GATA-1 in erythropoiesis via subunit Med1/TRAP220. *Proc Natl Acad Sci U S A* **103**, 18504-18509 (2006).

45. M. Levine, Transcriptional enhancers in animal development and evolution. *Curr Biol* **20**, R754-763 (2010).

46. C. Buecker, J. Wysocka, Enhancers as information integration hubs in development: lessons from genomics. *Trends Genet* **28**, 276-284 (2012).

47. T. Sawado, J. Halow, M. A. Bender, M. Groudine, The beta -globin locus control region (LCR) functions primarily by enhancing the transition from transcription initiation to elongation. *Genes Dev* **17**, 1009-1018 (2003).

48. S. H. Song *et al.*, Multiple functions of Ldb1 required for beta-globin activation during erythroid differentiation. *Blood* **116**, 2356-2364 (2010).

49. C. Lin, A. S. Garruss, Z. Luo, F. Guo, A. Shilatifard, The RNA Pol II elongation factor Ell3 marks enhancers in ES cells and primes future gene activation. *Cell* **152**, 144-156 (2013).

50. N. Dukler, B. Gulko, Y. F. Huang, A. Siepel, Is a super-enhancer greater than the sum of its parts? *Nat Genet* **49**, 2-3 (2016).

51. D. Hay *et al.*, Genetic dissection of the alpha-globin super-enhancer in vivo. *Nat Genet* **48**, 895-903 (2016).

52. H. Y. Shin *et al.*, Hierarchy within the mammary STAT5-driven Wap super-enhancer. *Nat Genet* **48**, 904-911 (2016).

53. M. Osterwalder *et al.*, Enhancer redundancy provides phenotypic robustness in mammalian development. *Nature* **554**, 239-243 (2018).

54. A. Mora, G. K. Sandve, O. S. Gabrielsen, R. Eskeland, In the loop: promoter-enhancer interactions and bioinformatics. *Brief Bioinform* **17**, 980-995 (2016).

55. M. C. King, A. C. Wilson, Evolution at two levels in humans and chimpanzees. *Science* **188**, 107-116 (1975).

56. A. Siepel, L. Arbiza, Cis-regulatory elements and human evolution. *Curr Opin Genet Dev* **29**, 81-89 (2014).