Data integration strategies for genome-wide discovery of enhancers

Iros Barozzi[a*#], Remo Monti[a*], Marco Osterwalder[a], Diane Dickel[a], Len Pennacchio[a,b], Axel Visel[a,b,c]

[a] *MS 84-171, Lawrence Berkeley National Laboratory, Berkeley, CA 94720, USA.*
[b] *U.S. Department of Energy Joint Genome Institute, Walnut Creek, CA 94598, USA.*
[c] *School of Natural Sciences, University of California, Merced, CA 95343, USA*

[#] *Correspondence: igbarozzi@lbl.gov*
* *Equal contributors*

Enhancers are short *cis*-regulatory DNA sequences that modulate the transcription of their target genes in a tissue- and condition-specific fashion. Indirect measurements suggest that hundreds of thousands of enhancers populate the non-coding portion of mammalian genomes, but only a few thousand of them have been validated for their activity *in vivo*. A recent deluge of genome-wide assays opened up opportunities to computationally model the potential of each non-coding region to act as enhancer. Here we introduce two approaches that tackle this problem from two complementary angles, using cardiac and limb enhancers as examples to provide proof of principle. The first one is an unbiased, integrative strategy, applied to scoring non-coding regions of the genome for their potential to be cardiac enhancers. The score considers >50 published ChIP-seq and DNase I accessibility datasets sampled across development and adulthood of both human and mouse. This compendium will be useful in many contexts, e.g. in human genetics studies of non-coding variation. The second approach combines the analysis of published datasets to computational prediction of transcription factor binding sites into a predictive model of enhancer activity in developing limbs. To this aim, it was trained on >200 hundreds enhancer sequences showing activity in limb at stage e11.5 of mouse development *in vivo*. Our model achieves better performance while offering a much better interpretability than previously published methods. The genome-wide predictions from both approaches are publicly available and can be easily consulted via UCSC genome browser. Both strategies can be generalized and applied to any mammalian system of interest in which a sufficient number of datasets is available. In particular, we are currently adapting both approaches for the identification of human craniofacial enhancers from a comprehensive set of chromatin profiling data across mouse and human craniofacial development that is being generated through FaceBase. We expect the resulting data sets to provide significantly improved maps of the genome-wide enhancer landscape underlying craniofacial development, with a multitude of applications in studies of craniofacial biology and disorders.