Due to rapid advances in next-generation sequencing technologies1, tens of thousands of disease-associated mutations2 and millions of single nucleotide variants3,4 have been identified in the human population. Previous studies claimed that “the majority (~93%) of disease- and trait-associated variants... lie within noncoding sequence”5. This drives an urgent need to develop high-throughput methods to sift through this deluge of sequence data to quickly determine the functional relevance of each noncoding variant. However, it has been shown that only a fraction of noncoding variants are functional, and among the functional variants, the majority show only modest effects6. Therefore, we need to develop highly quantitative assays that can examine a large number of variants. Furthermore, current computational methods (e.g., CADD7 and GWAVA8) widely applied in whole genome/exome studies are designed to prioritize “deleterious and pathogenic” variants, not designed to predict specific molecular phenotypes of these variants (i.e., their effects on enhancer activities).

To address this issue, we will establish a robust, high-throughput experimental pipeline with three novel technologies (***MegaMut***, ***STROBE-seq***, and ***CHAIN-seq***) to quickly clone >20,000 noncoding variants, and experimentally and quantitatively examine their impact on enhancer activity. All three assays offer distinct advantages over current state-of-art assays6,9-11, and together the overall pipeline will improve the throughput by at least an order of magnitude. Furthermore, the ultra high throughput of the new assays will allow us *for the first time* to establish a three-stage, iterative learning scheme (computational-experimental-integrated) with a real-time experimental parameter optimization strategy (i.e. in Year 3, we will carry out 6 rounds of experimental validation and refinement with ~1,000 selected variants cloned and examined per round; see **Fig. 1**). Finally, we will establish an integrated classifier model, ***ReEnAct***, through iterative learning to accurately predict functional molecular phenotypes for ***all*** noncoding variants. Our specific aims are:

**Aim 1. MegaMut: a massively parallel *en masse* site-directed mutagenesis pipeline.**

Building upon our experience in developing Clone-seq12, we will implement an *en masse* (“pooled”) site-directed mutagenesis pipeline, MegaMut, for introducing tens of thousands of mutations in thousands of enhancers of interest with *one and only one pre-determined mutation* on each DNA molecule. MegaMut is quite different from current assays. For instance, MPFD6 uses random mutagenesis to introduce mutations in three enhancers with on average three mutations per 100 bp on each molecule; MPRA9 use synthesized oligos and can only create clones <150 bp (vs. MegaMut with no length limit); PALS13 is designed to perform saturated mutagenesis for a given gene (two genes tested) and has higher unwanted mutations and WT background (compared to MegaMut). Clone-seq is currently one of the highest throughput site-directed mutagenesis pipeline. However, Clone-seq requires mutagenesis reactions be carried out individually and multiple single colonies be picked per reaction12, whereas MegaMut eliminate these steps by carrying out all mutagenesis reactions in a single pool, drastically increasing throughput and decreasing cost.

**Aim 2. STROBE-seq: a massively parallel quantitative assay to measure enhancer activity.**

Building upon our experience in developing GRO-seq (subsequently PRO-seq and GRO-cap)14-25, we will integrate GRO-seq with STARR-seq11 to quantitatively measure effects of variants on enhancer activity at a massive scale in a new assay called STROBE-seq. For our purpose, STARR-seq has several key limitations (see **b.2**). In particular, it requires creation of a complex library of unique but overlapping fragments for each enhancer tested11, and thus ***cannot*** be directly used to measure enhancer activities from a clone library of enhancer elements, where each element has one and only one clone with defined boundaries. Furthermore, DNA fragments >1 kb do not form clusters effectively in Illumina sequencing, which has been a huge roadblock in studying large enhancers by genomic assays. We designed a new strategy, CHAIN-seq, to sequence STROBE-seq libraries using paired barcodes, essentially eliminating any length limit of STROBE-seq.

**Aim 3. ReEnAct: an iterative learning framework to prioritize impactful noncoding variants.**

Building upon our experience in developing FunSeq (for prioritizing somatic mutations as potential cancer drivers), we will implement the ReEnAct classifier model to prioritize noncoding variants with a strong molecular phenotypic effect. In particular, **(a)** we will develop a set of DNA-level features associated with strong readout in our assays (e.g. TF motif breaking) and use it to consistently annotate all variants. **(b)** We will further prioritize variants that overlap with elements displaying strong allelic activity (ie demonstrated sensitivity to variants in various assays). **(c)** We will then use network connectivity from enhancer-gene linkages to further prioritize variants. **(d)** Finally, we will implement a iterative, three-stage computational-experimental-integrated learning scheme by coupling ReEnAct training inputs with MegaMut and STROBE-seq outputs. In particular, (i) we will clone and examine ~10,000 enhancer variants to initialize ReEnAct; (ii) we will carry out 6 rounds of real-time experimental optimization (~1,000 new variants/round); (iii) at the end, we will perform a final model assessment by examining ~5,000 new variants. In total, we will generate clones for ~3,000 WT enhancers and >20,000 variants, allowing for a comprehensive refinement and evaluation of ReEnAct.

**References Cited**

1. Snyder, M., Du, J. & Gerstein, M. Personal genome sequencing: current approaches and challenges. *Genes Dev* **24**, 423-431 (2010).

2. Stenson, P.D.*, et al.* The Human Gene Mutation Database: 2008 update. *Genome Med* **1**, 13 (2009).

3. Consortium, T.G.P. An integrated map of genetic variation from 1,092 human genomes. *Nature* **491**, 56-65 (2012).

4. Fu, W.*, et al.* Analysis of 6,515 exomes reveals the recent origin of most human protein-coding variants. *Nature* **493**, 216-220 (2013).

5. Maurano, M.T.*, et al.* Systematic localization of common disease-associated variation in regulatory DNA. *Science* **337**, 1190-1195 (2012).

6. Patwardhan, R.P.*, et al.* Massively parallel functional dissection of mammalian enhancers in vivo. *Nat Biotechnol* **30**, 265-270 (2012).

7. Kircher, M.*, et al.* A general framework for estimating the relative pathogenicity of human genetic variants. *Nat Genet* **46**, 310-315 (2014).

8. Ritchie, G.R., Dunham, I., Zeggini, E. & Flicek, P. Functional annotation of noncoding sequence variants. *Nat Methods* **11**, 294-296 (2014).

9. Melnikov, A.*, et al.* Systematic dissection and optimization of inducible enhancers in human cells using a massively parallel reporter assay. *Nat Biotechnol* **30**, 271-277 (2012).

10. Kheradpour, P.*, et al.* Systematic dissection of regulatory motifs in 2000 predicted human enhancers using a massively parallel reporter assay. *Genome Res* **23**, 800-811 (2013).

11. Arnold, C.D.*, et al.* Genome-wide quantitative enhancer activity maps identified by STARR-seq. *Science* **339**, 1074-1077 (2013).

12. Wei, X.*, et al.* A massively parallel pipeline to clone DNA variants and examine molecular phenotypes of human disease mutations. *PLoS Genet* **10**, e1004819 (2014).

13. Kitzman, J.O., Starita, L.M., Lo, R.S., Fields, S. & Shendure, J. Massively parallel single-amino-acid mutagenesis. *Nat Methods* **12**, 203-206, 204 p following 206 (2015).

14. Fuda, N.J.*, et al.* GAGA factor maintains nucleosome-free regions and has a role in RNA polymerase II recruitment to promoters. *PLoS Genet* **11**, e1005108 (2015).

15. Danko, C.G.*, et al.* Identification of active transcriptional regulatory elements from GRO-seq data. *Nat Methods* **12**, 433-438 (2015).

16. Core, L.J.*, et al.* Analysis of nascent RNA identifies a unified architecture of initiation regions at mammalian promoters and enhancers. *Nat Genet* **46**, 1311-1320 (2014).

17. Jonkers, I., Kwak, H. & Lis, J.T. Genome-wide dynamics of Pol II elongation and its interplay with promoter proximal pausing, chromatin, and exons. *Elife* **3**, e02407 (2014).

18. Kruesi, W.S., Core, L.J., Waters, C.T., Lis, J.T. & Meyer, B.J. Condensin controls recruitment of RNA polymerase II to achieve nematode X-chromosome dosage compensation. *Elife* **2**, e00808 (2013).

19. Saunders, A., Core, L.J., Sutcliffe, C., Lis, J.T. & Ashe, H.L. Extensive polymerase pausing during Drosophila axis patterning enables high-level and pliable transcription. *Genes Dev* **27**, 1146-1158 (2013).

20. Danko, C.G.*, et al.* Signaling pathways differentially affect RNA polymerase II initiation, pausing, and elongation rate in cells. *Mol Cell* **50**, 212-222 (2013).

21. Chopra, V.S.*, et al.* The polycomb group mutant esc leads to augmented levels of paused Pol II in the Drosophila embryo. *Mol Cell* **42**, 837-844 (2011).

22. Hah, N.*, et al.* A rapid, extensive, and transient transcriptional response to estrogen signaling in breast cancer cells. *Cell* **145**, 622-634 (2011).

23. Min, I.M.*, et al.* Regulating RNA polymerase pausing and transcription elongation in embryonic stem cells. *Genes Dev* **25**, 742-754 (2011).

24. Larschan, E.*, et al.* X chromosome dosage compensation via enhanced transcriptional elongation in Drosophila. *Nature* **471**, 115-118 (2011).

25. Core, L.J., Waterfall, J.J. & Lis, J.T. Nascent RNA sequencing reveals widespread pausing and divergent initiation at human promoters. *Science* **322**, 1845-1848 (2008).