Due to rapid advances in next-generation sequencing technologies1, tens of thousands of disease-associated mutations2 and millions of single nucleotide variants (SNVs)3,4 have been identified in the human population. 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 SNV. Previous studies, including our own, have shown that SNVs disrupting specific enhancer activities can lead to various disorders, including cancer. However, our preliminary results have shown that population SNVs are significantly less damaging than disease mutations, and current methods (e.g., CADD and GWAVA) widely applied in whole genome/exome studies are not designed to predict effects of SNVs on enhance activities.

To address this issue, we will establish a robust, high-throughput experimental pipeline with two novel technologies to quickly clone ~15,000 noncoding variants, and experimentally examine their impact on enhancer activity, providing *the first* systematic experimental measurements of biochemical functional impact of SNVs on the genomic scale. Furthermore, this will allow *for the first time* a thorough assessment of the relationships between these measured functional molecular phenotypes and inferred population-level fitness impact, as assessed by various population genetic metrics (e.g., allele frequency, *FST*, etc.). Finally, we will establish an integrated classifier through iterative learning to accurately predict functional molecular phenotypes for ***all*** noncoding variants. Overall, our proposed research will greatly improve *de novo* prediction accuracy of functional variants at the genomic level. Toward these goals, we propose three specific aims to develop two novel genome-scale technologies, both of which will improve the throughput at least 1-2 orders of magnitude higher than current state-of-art assays:

**Aim 1. MegaPFunkel-seq: a massively-parallel site-directed mutagenesis pipeline.**

Current protocols for cloning require the selection of individual colonies and subsequent sequencing of each colony using Sanger sequencing to find the correct clone5. The standard approach is both labor intensive and expensive, and does not scale well to high-throughput applications. Here we propose to implement an *en masse* (“pooled”) site-directed mutagenesis pipeline, MegaPFunkel-seq, for introducing mutations in transcription regulatory elements (TREs) of interest. We developed our MegaPFunkel-seq pipeline by incorporating mutagenesis megaprimers (electrochemically synthesized in large scale6) into a previously published method named PFunkel7. MegaPFunkel-seq is a high-throughput *site-directed* mutagenesis pipeline, so only *pre-determined mutation(s)* are introduced to targeted DNA sequences and each mutagenized DNA molecule will only contain those *pre-determined mutation(s)*, completely different from random mutagenesis assays8.

**Aim 2. STRO-seq: a massively-parallel quantitative assay for measuring enhancer activity.**

STARR-seq (self-transcribing active regulatory region-sequencing) is a recently established method that can identify enhancer elements genome-wide9. Briefly, short genomic fragments are cloned *en masse* into the 3’ untranslated region of a simple transcription unit between paired-end sequencing primers. After transfection of this fragment library into cells, enhancer activity is quantified by counting the number of unique fragments from a particular genomic locus that give rise to detectable mRNA. Importantly, STARR-seq does not quantify the enhancer activity of individual candidate fragments, but instead requires creation of a complex library of unique but overlapping fragments for each candidate region to be tested. It also requires that enhancer sequences can exist as stable mRNAs, and is thus confounded by post-transcriptional effects. Furthermore, >98% of sequencing reads are discarded because multiple mRNA molecules are often produced from a single unique DNA fragment (see Supplemental Figure 2E of Arnold et al9).

To circumvent these difficulties, we will develop a self-transcribing run-on sequencing (STRO-seq) protocol to allow direct quantification of enhancer activity for an individual enhancer sequence. After preparation of an enhancer library and transfection into cells, nascent RNA will be captured as in our established GRO-seq protocol. Importantly, candidate enhancer activity will be quantified as the number of nascent eRNAs produced per transfected plasmid. This approach offers many advantages: (1) reduced bias from post-transcriptional effects, (2) quantification of transcription driven by a specific enhancer fragment, and (3) more efficient use of sequencing reads. We anticipate that these improvements will significantly simplify high-throughput studies of candidate enhancer sequences, and increase assay sensitivity compared with STARR-seq. We will first generate clones for ~2,000 WT TREs and ~10,000 SNVs through MegaPFunkel-seq and examine their impact on enhancer activities using STRO-seq, providing training data for **Aim 3**.

**Aim 3. eleVAR: a computationally-experimentally-integrated iterative learning framework for prioritizing impactful non-coding variants.**

Our eleVAR pipeline will build upon the FunSeq approach, which we developed earlier, to prioritize rare germline variants that occur within genomic regions under negative selection within the human population. **(a)** We will enlarge this approach to consistently annotate all the existing set of DNA-level features. This will include TF binding sites, as well as non-coding RNA features such as RNA-binding-protein sites and structured regions. **(b)** We will further prioritize variants that overlap with genomic elements that display strong allelic activity. This will require a large-scale calculation to identify allelic elements. **(c)** We will then use network connectivity from predicted enhancer/promoter-gene linkages, microRNA targeting, and other sources to prioritize variants at hubs and bottlenecks. **(d)** Finally we will use an entropy-based integrated scoring scheme to combine this diverse set of features into a score for each variant in the genome.

We will perform three rounds of iterative validation and learning of parameter weights to improve eleVAR. **(a)** In each of the first two rounds, we will choose 400 genomic elements, and use our newly-developed massively-parallel MegaPFunkel-seq pipeline to generate five variants, two highly prioritized (predicted to be deleterious by eleVAR), one with a medium score, and two with lower scores (2,000 total variants/round). We will assess impact of variants on gene regulation using STRO-seq, comparing wild type and predicted high and low impact variants. **(b)** In the final round, purely for assessment, we will choose another 200 genomic elements and five variants with high, medium, and low scores on each element and generate clones using MegaPFunkel-seq. In total, we generate clones for ~1,000 WT genomic elements and ~5,000 variants, which will allow for a comprehensive evaluation of the eleVAR performance.

**Successful completion of the proposed project will generate a number of highly impactful resources:**

1. A clone library of ~3,000 WT TREs and ~15,000 noncoding variants and a publicly accessible database of their effects on gene regulation, which will fuel hypothesis-driven studies around the world.
2. A high-throughput technology to rapidly clone and examine a large number of noncoding SNVs, the general scheme of which can be readily expanded to include other methods.
3. One robust machine-learning pipeline to predict impacts on gene regulation for ***ALL*** noncoding SNVs in the human population.

**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. Suzuki, Y.*, et al.* A novel high-throughput (HTP) cloning strategy for site-directed designed chimeragenesis and mutation using the Gateway cloning system. *Nucleic Acids Res* **33**, e109 (2005).

6. Maurer, K.*, et al.* Electrochemically generated acid and its containment to 100 micron reaction areas for the production of DNA microarrays. *PLoS One* **1**, e34 (2006).

7. Firnberg, E. & Ostermeier, M. PFunkel: efficient, expansive, user-defined mutagenesis. *PLoS One* **7**, e52031 (2012).

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

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