a. SIGNIFICANCE

a.1. Non-coding variants are significant for disease but less well-studied than coding ones

Numerous studies have been conducted on the mutations that lie in coding regions[1-4]. Not as much has been done on non-coding ones. However, several initial studies suggest that variants in non-coding regions can significantly influence an organism's phenotype[5,6], and they are often implicated in diseases[7-9]. Many non-coding variants impact regulatory elements. Such variation in the human genome can modulate gene expression[10], and changes in this expression have been implicated in cancer and other diseases[11-16].

a.2. Recent progress in annotating non-coding regions of the genome provides new opportunities for variant interpretation

Annotating non-coding regions is essential for investigating genome evolution[17], understanding important biological functions (including gene regulation and RNA processing)[18], and for elucidating how SNPs and structural variation may influence disease[19]. The Encyclopedia of DNA Elements (ENCODE) and the model organism ENCODE (modENCODE) Project provide extensive genomic annotation of human[20], drosophila[21] and *C. elegans*[22] genomes. Furthermore, the functional landscape of regulatory variations in the human genome has been investigated by large-scale mRNA and miRNA sequencing (Fig 1)[23-25]. Similar efforts have also been directed toward annotating human epigenomic data[26], as well as understanding the influence of genomic variation on the gene expression profiles[27]. These Expression Quantitative Trait Loci (eQTL) can further be utilized to investigate underlying disease mechanisms[28].

a.3. Molecular phenotypes help understand epistasis and identify actionable drug targets

In this proposal, we are interested in the molecular phenotypes of SNVs, because all genetic lesions leading to organismal disease phenotypes or affecting overall fitness must have an underlying molecular basis[29]. We fully acknowledge that some SNVs with significant molecular phenotypes will not lead to any disease or traits. This can happen due to epistasis between TREs/genes with redundant functions. In fact, epistasis is a major roadblock to studying these TREs/genes genetically at the cellular and organismal level[7,30]. Our approach bypasses this limitation by directly examining the effects of variants on gene regulation at the molecular level. Even though mutations on one of these functionally redundant TREs/genes may not cause a disease phenotype, if these mutations disrupt the function of the corresponding TRE/gene (which can be measured by the methods proposed here), they will significantly increase the predisposing risk for disease[7,31,32]. Furthermore, accurate measurement of each SNV's impact on gene regulation are is essential for generating concrete hypotheses about disease etiology based on molecular mechanisms[7,33,34]. Such specific predictions are also vital for selecting actionable drug targets[35,36] and ultimately for making tailored therapeutic decisions[37,38], which are all crucial for the Precision Medicine Initiative[39].

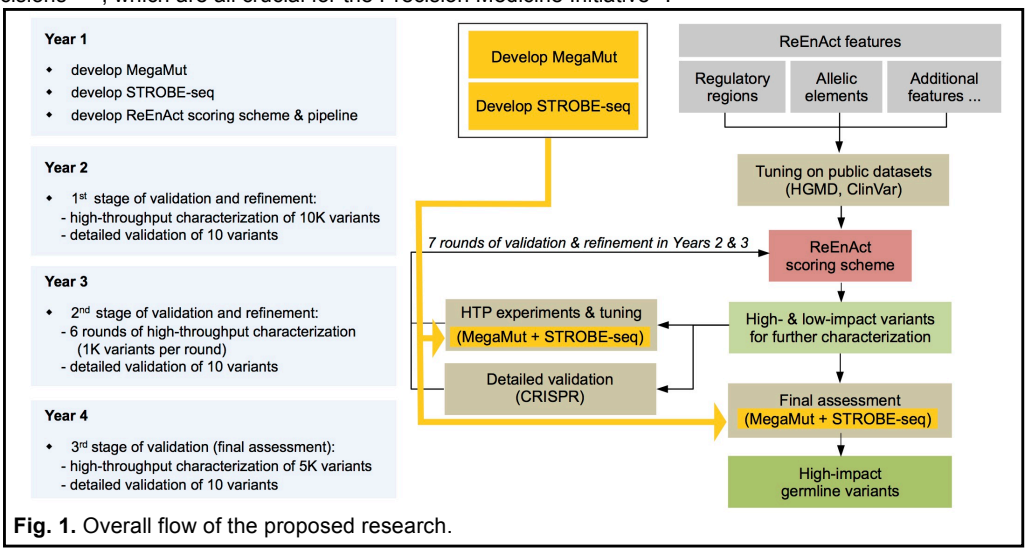

Fig. 1. Overall flow of the proposed research.

b. INNOVATION

b.1. MegaMut: an *en masse* site-directed mutagenesis pipeline

MegaMut allows massively-parallel site-directed mutagenesis to generate ***one and only one specific*** mutation per DNA molecule for ***thousands*** of enhancers. It is distinct from what was used in the most closely related previous studies: the massively parallel functional dissection (MSFD) approach[40] and the massively parallel reporter assay (MPRA)[41,42]. For MSFD, random mutagenesis was used to generate mutations for ***three*** mammalian enhancers[40]. In random mutagenesis, control of the number of mutations generated on each DNA sequence is impossible. To improve coverage, most random mutagenesis pipelines generate on average two or more mutations on each DNA molecule[43] and in the MSFD approach "each synthetic enhancer molecule contained, on average, three mutations per 100 bp, randomly distributed along its length"[40]. This prevents assessment of the functional impact of each individual mutation.

For MPRA, microarray-based oligo synthesis was used to generate mutations[41,42]. Although specific mutations can be generated in this approach, the length limit is <150 bp. However, depending on the definition used, mammalian enhancers can be 1 kb or longer[40]. We have used our MegaMut to generate hundreds of mutant TRE clones with an average length of ~500 bp. We will have no problem cloning enhancers (up to 4kb) and their mutations in their entirety.

Furthermore, MegaMut has many advantages comparing with PALS (Programmed Allelic Series)[44], a recently developed site-directed mutagenesis technique. While MegaMut and PALS share similar goals, PALS has only been used to simultaneously generate mutations in one target gene at a time (two genes total). Our MegaMut is capable of introducing multiple mutations to hundreds of target genes/TREs in one reaction. More importantly, MegaMut demonstrates a much lower background, i.e. much fewer clones with undesired mutations or with WT sequences. In addition, MegaMut allows us to evaluate the mutagenesis performance of every batch of experiment before sequencing while PALS relies solely on DNA sequencing to examine the performance. Last but not least, PALS uses subassembly[44,45] strategy to sequence the mutagenesis products by Illumina and have difficulty sequencing DNA elements longer than ~1.5 kb. However, our MegaMut + CHAIN-seq pipeline has no problem handling full-length sequences of much longer gene/TRE elements.

Finally, our newly-established Clone-seq is currently one of the highest throughput site-directed mutagenesis pipelines, through which we have successfully generated 1,034 mutant clones for 223 different genes[46]. However, Clone-seq requires that an individual mutagenesis PCR reaction be performed and multiple single colonies be picked for each mutation, whereas MegaMut allows all mutagenesis reactions are carried out *en masse* in one pool. Therefore, MegaMut offers distinct advantages and enables us (1) to order pooled mutagenesis primers (only one primer for each mutation, instead of a pair of primers in Clone-seq) through microarray-based oligo synthesis, reducing primer cost by >50×; (2) to perform pooled PCR reactions, instead of individual mutagenesis PCRs in Clone-seq; and (3) to avoid colony picking, the rate-limiting step in Clone-seq. Overall, our novel MegaMut pipeline significantly reduces the cost of large-scale site-directed mutagenesis and improve the throughput at least one to two orders of magnitude.

b.2. STROBE-seq: *S*elf-*t*ranscribing nuclear *r*un-on with paired *b*arcodes and *seq*uencing assay

STARR-seq (self-transcribing active regulatory region-sequencing) is a recently established method that can identify enhancer elements genome-wide[47]. 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. Thus it ***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. 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 Sup. Fig. 2E of Arnold et al[47]).

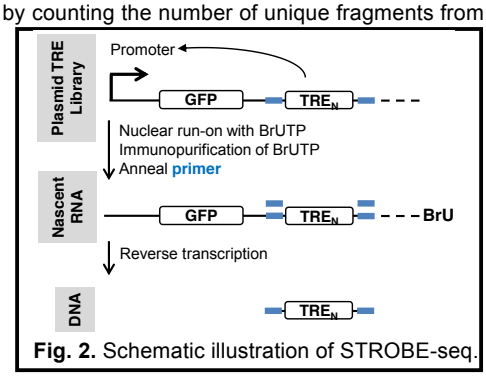

Fig. 2. Schematic illustration of STROBE-seq.

To circumvent these difficulties, we will develop a self-transcribing run-on with paired barcodes and sequencing (STROBE-seq) protocol to allow direct quantification of enhancer activity for an individual enhancer sequence (**Fig. 2**). 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, (3) more efficient use of sequencing reads, and (4) ability to sequence large enhancer elements through CHAIN-seq (**b.3**). These improvements will significantly simplify high-throughput studies of candidate enhancer sequences, and increase assay sensitivity compared with STARR-seq.

b.3. CHAIN-seq: chains of paired barcodes for sequencing of large targets

A major limitation in many studies of TREs has been the difficulty of sequencing elements larger than 1 kb using Illumina. Notably, many large so-called "super" or "stretch" enhancers, whose size is often >3 kb[48], have recently been found to be particularly relevant throughout development and carcinogenesis[49]. To study these elements, we will develop a novel strategy, called CHAIN-seq, to comprehensively sequence large DNA elements by implementing a customized transposase reaction to insert paired barcodes into the sequencing target (**Fig. 2**). The key idea of CHAIN-seq is to perform two separate Illumina runs on the aliquots of the same sample: one regular single-end Illumina run to sequence all the barcode-barcode pairs; another regular paired-end Illumina run to sequence the target fragments with barcodes. After the second run, one can easily re-assemble the large target sequence by connecting all the neighboring reads through the barcode-barcode pairs determined by the first Illumina run. Using this strategy, we will be able to clone and sequence all potential enhancer elements through STROBE-seq without size limit. In fact, CHAIN-seq strategy could be useful for a wide variety of sequencing applications. For example, it can be applied to *de novo* genome sequencing and significantly reduce the difficulty of sequence assembly.

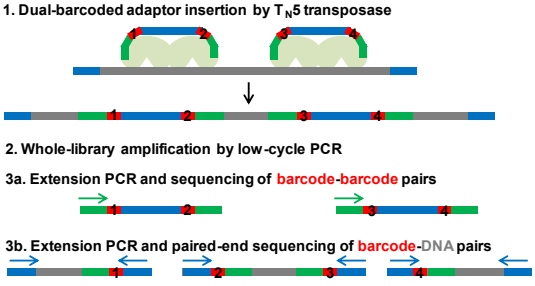

Fig. 3. Schematic illustration of CHAIN-seq.

b.4. ReEnAct: an innovative iterative learning scheme utilizing allelic activity with real-time experiential parameter tuning

Previous studies have identified specific variants using allele-specific activity[50,51]. However, there has not been a scheme that allows us to prioritize variants based on this. In the proposed work, we will prioritize variants based on their presence within allelic elements or regions of the genome. Another innovative aspect of this proposal is the parameter weighting scheme and iterative tuning. In the first iteration, we will implement a weighted scoring scheme by assigning weights to various features based on publicly available polymorphism data. Each variant will be assigned a weighted score based on the weight of individual features associated with that particular variant. In the subsequent iterations of this workflow, we will apply a Bayesian learning strategy to tune weights based on experimental observations on >20,000 SNPs within ~3,000 enhancer elements. Because of the high throughput of MegaMut and STROBE-seq, in Year 3 we will *for the first time* perform real-time experimental parameter tuning of ReEnAct by cloning and experimentally examining ~1,000 SNPs every two months (6 rounds of iterative tuning in one year). We believe this type of truly computational-experimental-integrated machine learning scheme has never been implemented before, and it is only possible because of the development of MegaMut and STROBE-seq with their ultra high throughput.

OVERLAP?

c. APPROACH

c.1. Specific Aim 1. MegaMut: a massively-parallel site-directed mutagenesis pipeline.

In this aim, we will build upon our experience in developing the Clone-seq pipeline to establish the MegaMut pipeline with a much higher throughput (at least one to two orders of magnitude higher).

c.1.1. Preliminary Studies

c.1.1.1. Clone-seq: a massively-parallel cloning pipeline. Current protocols for cloning require the selection of individual colonies and subsequent sequencing of each colony using Sanger sequencing to find the correct clone[52]. The standard approach is both labor intensive and expensive, and does not scale well to high-

throughput applications. *In Clone-seq[46], we implement a "smart-pooling" strategy to put single colonies of each cloning attempt into one pool and combine multiple pools through multiplexing for one Illumina sequencing run such that we can distinguish sequencing reads for each colony of each clone computationally afterwards.* We have successfully generated 1,034 clones in an optimized high-throughput fashion[46]. Using our customized variant calling software, we identified correct clones free of any other unwanted mutations introduced during PCR. We achieved a conservative estimate in cost-savings of at least 10-fold over conventional cloning[46], which can be further improved with implementation of newer sequencing platforms (e.g., NextSeq 500).

Clone-seq is very versatile. It can be used to generate wild type gene/TRE clones or specific mutant clones[46]. We have successfully generate ~800 wild type TRE (enhancers and promoters) clones. Our results confirm that Clone-seq can successfully generate clones for the ~3,000 wild type enhancers within the proposed time frame.

c.1.1.2 MegaMut, an en masse ("pooled") site-directed mutagenesis pipeline. Owing to the nature of our STROBE-seq assay, there is no need to generate separate clones for each individual mutant TRE. Here we propose to implement an *en masse* ("pooled") site-directed mutagenesis pipeline, MegaMut (**Fig. 4**), for introducing mutations in TREs of interest. We developed our MegaMut pipeline by incorporating mutagenesis megaprimers (electrochemically synthesized in large scale[53]) into a previously published method named PFunkel[54]. MegaMut 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)***. As a preliminary result, we generated 212 specific mutations for 53 different WT DNA clones with minimal undesired mutations in one MegaMut reaction and ~20% WT background, consistent with previous PFunkel publications[54,55]. In another preliminary study, we spiked in a particular megaprimer at 1/1,000 concentration compared to other megaprimers in the pool and were able to detect successful mutagenesis products (**Fig. 7**). Therefore, we are highly confident that we will be able to optimize the MegaMut pipeline to perform >1,000 mutagenesis reactions in one pool, and successfully generate >20,000 noncoding mutation clones through pooling and multiplexing as proposed.

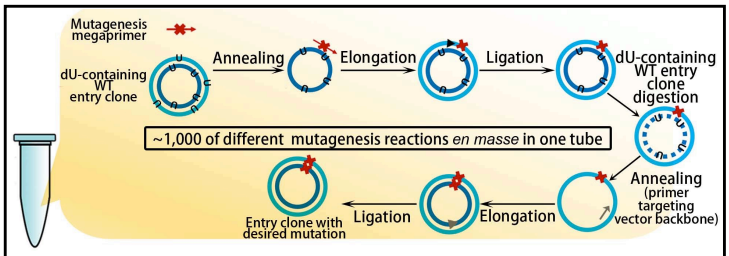

Fig. 4. Schematic illustration of MegaMut.

c.1.2. Research Design

c.1.2.1. Developing the MegaMut assay. Megaprimer design, synthesis, and trimming. The mutagenesis primers ("megaprimers") are ssDNA oligos electrochemically synthesized on a programmable DNA microarray (CustomArrray) and released into solution. The total length of each megaprimer will be 160nt, which includes 120nt of template (TRE)-priming region flanked by two 20nt adaptors and is 5'-phosphorylated (**Fig. 5**). The template-priming region contains the target mutation and is designed in batch with MutPrimer[46]. The 5'-adaptor contains different pre-determined barcodes and always ends with a thymine. The 3'-adaptor always begins with an adenine. PCR with amplification primers complementary to the adaptors will be used to amplify the megaprimers. The special design/modification of megaprimers and amplification primers allows for the selective amplification of subgroups of primers of interest for "smart pooling" (see **c.2.2.4**), as well as easy removal of both adaptors before the mutagenesis reaction (**Fig. 5**). In addition, a primer complementary to the entry clone backbone sequence (P_{BB}) will be synthesized with 5'-phosphorylation.

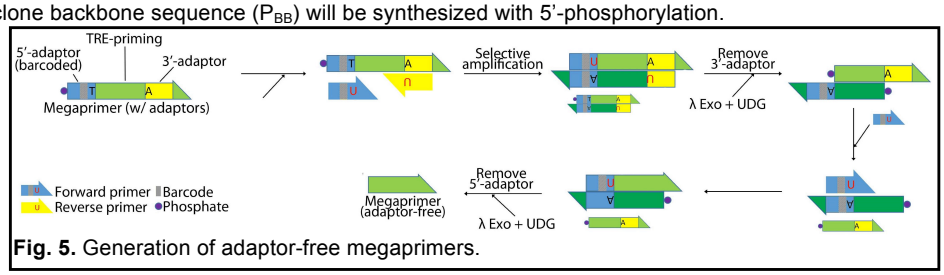

Fig. 5. Generation of adaptor-free megaprimers.

Preparation of WT TRE templates for mutagenesis. To prepare uracil-containing WT TRE templates, entry clones carrying TREs of interest will be propagated in CJ236 *E. coli*.

En masse mutagenesis reactions. Uracil-containing WT TRE entry clones and their corresponding mutagenesis megaprimers (adaptors removed) will be combined in a single reaction volume. In addition, the reaction mixture also contains PfuTurbo Cx hotstart DNA polymerase, dNTPs, Taq ligase, DTT, NAD+, and buffer. A two-step PCR will be performed with 12 cycles with megaprimers being gradually added to the reaction mixture (preceding Cycle 1, 5, and 9). The PCR reaction will be followed by a ligation step before uracil-DNA glycosylase (UDG) and Exonuclease III are added to remove the WT uracil-containing TRE templates while leaving the circular, mutation-bearing, and uracil-free ssDNA intact. After heat inactivation of the nuclease, the primer complementary to the entry clone backbone sequence (P_{BB}) will be added to the reaction mixture and the complementary strand will be synthesized by one cycle of PCR. A ligation step will be performed afterwards to generate circular, uracil-free dsDNA with desired mutations, which will be transformed into TOP10 *E. coli* through electroporation. All transformants will be plated on LB-agar plates with appropriate antibiotics and all yielded *E. coli* colonies will be harvested and propagated in liquid culture for plasmid extraction.

c.1.2.2. Optimizing the MegaMut protocol. Reduction of WT background in MegaMut products by pretreating uracil-containing WT TRE entry clone plasmids. After being extracted from CJ236 *E. coli*, the uracil-containing WT entry clone plasmids will be subjected to Exonuclease V digestion to remove most fragmented plasmids and bacterial genomic DNA.

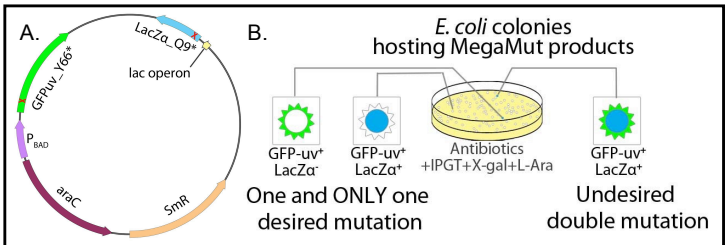

Fig. 6. (A) pUC-Double-Null vector. (B) MegaMut performance can be easily examined without Illumina sequencing

"Spiking-in" strategy for sequencing-free MegaMut performance evaluation. For pre-Illumina sequencing performance evaluation of MegaMut reaction, we will use a "spiking-in" strategy. We engineered a control plasmid (pUC-Double-Null, **Fig. 6A**) carrying a null-mutated lacZα CDS and a null-mutated GFP-uv CDS. Two megaprimers (LacZα_RES and GFPuv_RES) that rescue either null mutation respectively were designed and tested. pUC-Double-Null and the two megaprimers will be spiked in each MegaMut reaction with in proportion to the WT TRE plasmids and their corresponding megaprimers. The transformants will be plated on LB agar plates supplemented with antibiotics, L-arabinose, X-gal, IPTG. By simply examining the colonies on LB plates with our bare eyes under regular ambient light and a UV lamp, we can easily identify *E. coli* colonies that are (1) blue and non-fluorescent, (2) white and green-fluorescent and (3) blue and green-fluorescent. Quantification of those colonies will give us an estimation of the frequencies of desired mutagenesis products (1 & 2) and undesired products from double-annealing (3), as depicted in **Fig. 6B**.

Optimization of different parameters of MegaMut with the lacZα/GFP-uv assay for higher throughput and robust performance. To further increase the throughput of MegaMut, i.e. the number of unique mutant clones per reaction volume, we developed a LacZα/GFP-uv assay for estimating the maximal capacity of MegaMut as well as for tuning various parameters in the protocol. In lacZα/GFP-uv assay, the two template plasmids are pUC-LacZα-Null and pUC-GFPuv-Null. The two plasmids were generated by cloning either a lacZα CDS with a null mutation (pUC-LacZα-Null) or a GFP-uv CDS with a null mutation (pUC-GFPuv-Null) into the same backbone. MegaMut reaction will be carried out with these two plasmids and megaprimers LacZα_RES and GFPuv_RES. By reducing the percentage of pUC-LacZα-Null and LacZα_RES in the reaction volume (100%, 10%, 1%, 0.1%, 0.05%), we find that MegaMut can robustly generate ~1,000 mutations in one reaction; even at 1:2,000 dilution, we still consistently detected blue colonies (**Fig. 7**). Furthermore, with different combinations of pUC-LacZα-Null and pUC-GFPuv-Null, this lacZα/GFP-uv assay will be used to adjusting other reaction parameters: template/megaprimer ratio, number of PCR cycles, and so on.

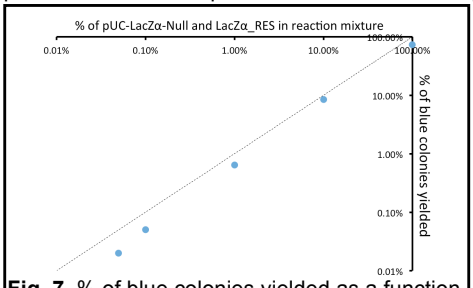

Fig. 7. % of blue colonies yielded as a function of % of pUC-LacZα-Null & LacZα_RES.

c.1.2.3. High-throughput cloning of ~3,000 WT TREs using Clone-seq. Sequence-specific forward and reverse primers containing attB1 and attB2 sequences, respectively, will be synthesized in bulk as "Trumer Oligo" plates by Eurofins Genomics[46]. Using human genomic DNA as template, the selected WT TREs will be PCR amplified in 96-well format with high-fidelity Phusion DNA polymerase. We will perform large-scale Gateway BP reactions to clone each PCR product into pDONR223 vector. Entry clones containing the intended TREs will be identified through our Clone-seq protocol[46]. The verified entry clones will be used for the site-directed mutagenesis by MegaMut. These WT entry clones will also be subjected to Gateway LR reaction to transfer TREs in the entry vector to our modified pDEST-hSTROBE destination vectors via recombination. The resulting expression clones will be pooled, maxipreped, and subjected to STROBE-seq analysis in K562 cell line to serve as the baseline control.

c.1.2.4 Potential pitfalls and alternative approaches. All current *en masse* mutagenesis techniques tend to have a relatively high WT allele background[44] and similar problems was also observed in the original Kunkel method[56] and its derivative, e.g. PFunkel[54,55]. In Kunkel-based method, the WT background is considered to result from WT plasmid fragments priming to circular WT templates during the first round of PCR reaction. To further minimizing any residual WT plasmid fragments from serving as "primers" in MegaMut reactions, we will perform one round of pre-mutagenesis extension reaction with the presence of Exonuclease V-treated plasmids, DNA polymerase, and ddNTPs (**Fig. 8**). This allows the incorporation of ddNTPs at the end of any WT plasmid fragment and prevent further elongation during subsequent mutagenesis PCR cycles. In addition, we use megaprimers with 120nt priming region instead of regular mutagenesis primers with ~35nt priming region. Such long priming region allows us to use a much higher annealing temperature and thus favors annealing of megaprimers over smaller WT plasmid fragments.

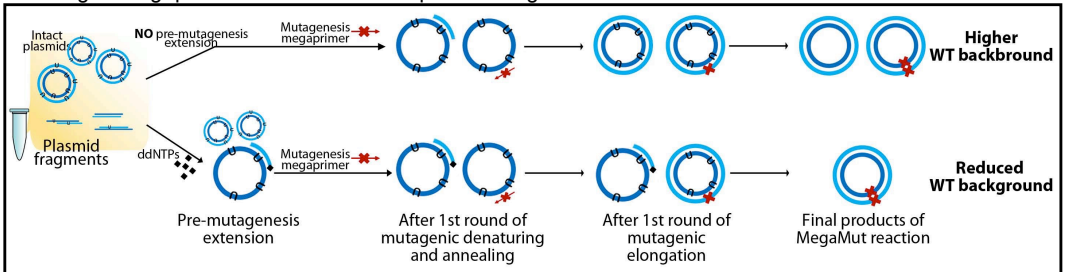

Fig. 8. Pre-mutagenesis extension with ddNTPs will further reduce WT background in MegaMut reactions.

Furthermore, based on our preliminary results, we have found that the number of PCR cycles directly affects the level of WT background: low PCR cycles lead to significantly high WT background; whereas high PCR cycles cause higher unwanted mutations. Our various preliminary experiments confirm that the use of 12 cycles as described in **c.1.2.1** is a good starting point. However, we will further optimize this, together with different template/megaprimer ratios, annealing temperatures, extension time, and so on.

Although unexpected additional mutations (mutations not included in the megaprimers) were shown to be rare in our preliminary results, in agreement with previous publications using the PFunkel method[54,55], we are still prepared to tackle the problem by controlling the three major sources of undesired mutations: (1) errors introduced by DNA polymerase during PCR, (2) simultaneous annealing and extension of multiple different megaprimers to the same WT template, and (3) non-specific priming. Using PfuTurbo Cx Hotstart DNA polymerase for mutagenesis PCR reaction will help minimize the DNA replication error, as it is a high-fidelity DNA polymerase that efficiently reads through a uracil base. Using long megaprimers and high annealing temperature will help reduce undesired mutations from both source (2) and (3). Last but not least, we gradually add small quantities of megaprimers to the MegaMut reaction mixture at different cycles of the PCR reaction, maintaining a high template/megaprimer ratio during the whole PCR process, which will further discourage different megaprimers from simultaneously annealing to the same template. If simultaneous annealing turns out to be more of a concern for certain long target TREs, we will use a "smart pooling" strategy: each megaprimer pool (associated with one barcode) will be amplified selectively with corresponding amplification primers and added to individual mutagenesis reactions to keep the one-to-one megaprimer-to-TRE relationship in each reaction, eliminating undesired simultaneous priming.

Previous studies all relied on Illumina sequencing to evaluate the performance of *en masse* mutagenesis techniques[44,54,55]. Such strategy is not only expensive but also time- and labor-consuming. Here we developed

a "spiking-in" strategy, allowing for pre-Illumina sequencing evaluation of MegaMut performance with ***NO*** additional experimental procedure or cost (just need to count colonies under ambient light and UV light).

Last but not least, if we have difficulty generating certain mutant alleles with MegaMut, we will use our robust Clone-seq pipeline to generate these mutant clones. Our Clone-seq pipeline has been optimized and used to generate >2,000 mutant clones in our lab and is more than capable of generating >6,000 TRE mutants within the proposed timeframe and budget.

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

With the WT and mutant enhancer clones generated in **Aim 1**, we will develop the STROBE-seq assay to quantitatively measure the enhancer activity of each clone and detect mutations that significantly change the enhance activities over the corresponding WT clones.

c.2.1. Preliminary Studies

c.2.2.1. Identification of TREs by GRO-seq and GRO-cap experiments. Global Run-On Sequencing (GRO-seq), identifies the location, orientation, and amount of all transcriptionally engaged RNA polymerases genome-wide, revealing the transcriptional landscape with an order of magnitude greater sensitivity than Pol II ChIP-seq methods. A derivative method, Global Run-On Cap sequencing (GRO-cap) maps the position of the transcription start-sites (TSSs) with base-pair resolution. The Lis lab has already performed genome-wide GRO-seq and GRO-cap experiments in GM12878 and K562 cells to identify TREs (both promoters and enhancers) defined by characteristic bidirectional TSSs[57].

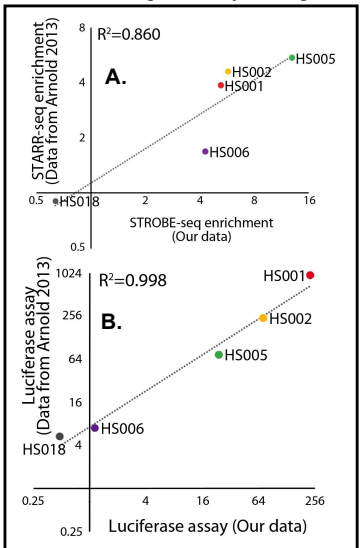

Fig. 9. Our results agree well with published data.

c.2.1.1. Modified Gateway-compatible STROBE-seq vectors. To make the STARR-seq compatible with our high-throughput cloning/mutagenesis pipeline, we modified the original STARR-seq vector by substituting the flanking homology arms with a Gateway cassette (attR1-R2) and retaining the Developmental Core Promoter (dCP). Our modified vector (called pDEST-hSTROBE-dCP) behaves like the original vector in transfection assays. We generated entry clones carrying four genomic DNA fragments (HS001, 002, 005, 006) that showed enhancer activity and one (HS018) that did not as measured by STARR-seq previously[47]. We cloned the five fragments in pDEST-hSTROBE-dCP by Gateway LR reaction, transfected them into HeLa cells, and quantified transcripts from each by qRT-PCR. Additionally, all five fragments were also cloned into pGL4.23-DEST-dCP vector and their enhancer activity was also confirmed by the dual luciferase assay as described in **c.1.2.3**. Both experiments (**Fig. 3**) successfully recapitulated the data published in the original STARR-seq paper[47]. Thus, the Gateway-compatible STROBE-seq vector is compatible with our high-throughput cloning/mutagenesis pipeline, and provides reliable quantification of the enhancer activity of target DNA fragments. To ensure coverage of the main classes of enhancers, we will use STROBE-seq vectors representing the two major classes of core promoters[58]: one that is responsive to developmental enhancers (pDEST-hSTROBE-dCP) and one responsive to housekeeping enhancers (pDEST-hSTROBE-hkCP).

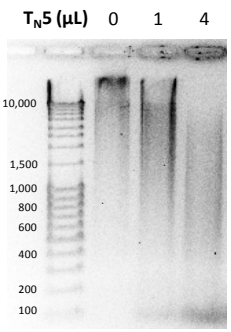

Fig. 10. Our customized tagmentation reaction.

c.1.2.3. Purification of transposase for use with custom oligonucleotide adaptors. To develop CHAIN-seq, it will be necessary to load transposase with custom oligonucleotide adaptor sequences containing random barcode pairs. To this end, we followed a recently published protocol[59] to express and purify T_N5 transposase in BL21-DE3 *E. coli* cells. To assay the activity of our enzyme preparation, we pre-loaded transposase with the customized Illumina sequencing adaptors and incubated this mixture with high-molecular weight human genomic DNA (**Fig. 10**). As expected, we observed dose-dependent tagmentation of the target DNA, validating our ability to tagment target DNA with customized adaptor sequences.

c.2.2. Research Design

c.2.2.1. Developing the STROBE-seq assay. STROBE-seq library transfection. The WT and mutant entry clones generated in **Aim 1** will be cloned into both pDEST-hSTROBE-dCP and pDEST-hSTROBE-hkCP vectors by Gateway LR reaction. Each resulting STROBE-seq library will be transfected into K562 cells by electroporation.

Self-transcribing run-on reaction. Nuclei will be isolated at 24h post-transfection. Briefly, cells are rinsed with ice cold PBS and lysed with a dounce homogenizer. Run-on reactions using 5-bromouridine 5′-triphosphate (BrUTP) will be performed, as we have previously described[60]. Nascent RNAs will be captured by immunopurification of BrUTP, and plasmids will be harvested from the flow-through of the immunopurification using EZNA Plasmid DNA columns (as Input control). Nascent RNAs will be purified by phenol-chloroform extraction and ethanol precipitation. Nuclear run-ons have been successfully performed on plasmids transfected into human cells[61], further supporting the feasibility of this approach.

Illumina sequencing library preparation. Nascent RNAs captured from the run-on reaction will be reverse transcribed from priming sites specific to pDEST-hSTROBE vectors. PCR-amplification will be performed as previously described[47]. Briefly, 1st-strand cDNA will be synthesized with a vector backbone-specific primer downstream of the cloning site (**Fig. 2**). This primer will contain the Illumina adaptor sequence, and a random barcode to enable removal of PCR duplicates. We will use multiplex-barcoded PCR primers to enable pooling of multiple samples in a single high-throughput sequencing lane.

c.2.2.2. Use of CHAIN-seq for analysis of large enhancers with STROBE-seq. An 89 nt oligonucleotide has been designed to form homodimers, allowing extension to form a full, dual-barcoded palindromic adaptor (**Fig 4**). The 134 bp adaptor will be size-selected on a 1.5% agarose gel and purified with a QIAGen Gel Extraction Kit. Full-length adaptors will be loaded onto purified T_N5 transposase and incubated with target DNA as previously describe[59]. With transposase binding sequences at the 3' end of both strands, the adaptor will be inserted by transposase dimers without creation of a double-strand break. To enable analysis of WT and mutated enhancers up to 4 kb long, we will apply CHAIN-seq (**Fig. 3**) to the transfected plasmids (Input) and PCR-barcoded reverse transcription products. Importantly, dosage of the transposase will be necessary to determine a working concentration leading to an integration density of approximately one every 400 bp of target DNA. This is well above the reported maximum density of ~30 bp[59].

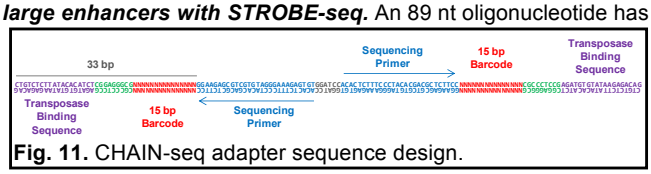

Fig. 11. CHAIN-seq adapter sequence design.

Following insertion of the CHAIN-seq adaptor, the resulting DNA libraries will be duplicated by 4 cycles of PCR using backbone-specific primers, purified, and split into two separate aliquots. One aliquot will be amplified with primers targeting the Transposase Binding Sequence (TBS), resulting in amplicons containing barcode pairs (**Fig. 3 3a**). These primers will also contain the necessary Illumina flow-cell adaptor sequences. This pool will be subjected to high-throughput sequencing using the TBS as a priming site.

The other aliquot of DNA will be amplified with primers targeting the central adaptor sequences, creating amplicons containing sample DNA and 2 random barcodes (**Fig. 3 3b**). Paired end sequencing of these amplicons (HiSeq 2×250 or NextSeq 2×150) will reveal the target DNA sequence flanking the integration site and the unique barcode associated with each integration site. Analysis of these sequences, combined with the barcode pairings discovered from the first aliquot, will enable reconstruction of large contiguous sequences with sufficient coverage for accurate variant calling of large enhancer libraries.

c.2.2.3. Measuring effects of ~20,000 TRE mutations on enhancer activity by STROBE-seq. The mutant TRE entry clones will be subcloned into pDEST-STROBE-dCP and pDEST-STROBE-hkCP vectors by high-throughput Gateway LR reactions. The enhancer activities of the WT and mutant TREs will be carefully examined by our STROBE-seq experiments as described above. A subset of mutations that show significant effects on enhancer activity (either up or down) will be further validated by high-throughput Luciferase assays as described in **c.2.2.4**.

c.2.2.4. High-throughput dual luciferase assay confirmation of STROBE-seq results. The luciferase reporter vector pGL4.23 (Promega) was modified into two Gateway compatible vectors, pGL4.23-DEST-dCP and pGL4.23-DEST-hkCP. These vectors contain a Gateway cassette upstream of the corresponding core promoter (dCP and hkCP) followed by a luc2 (synthetic firefly luciferase) reporter gene. Based on our STROBE-seq results with two different core promoters, TREs of interest will be LR-cloned into these reporter vectors accordingly. pGL4.75 vector (Promega), which contains a CMV enhancer/promoter and a downstream hRluc (synthetic Renilla luciferase) gene, is used as transfection control. TRE-containing reporter vector and

control vector will be co-transfected into K562 cells. The enhancer activity of TREs as indicated by the intensity of bioluminescence will be measured by with Dual-Glo luciferase assay system.

c.2.2.5. Detailed validation of selected noncoding variants at native loci in CRISPR knock-in cell lines using high-resolution and high-sensitivity assays: PRO-seq, 4C, and ChIP. Although STROBE-seq and the related STARR-seq are powerful high-throughput methods that enable one to test potential enhancer activity of DNA fragments of interest, we are well aware of its limitations. Most importantly, the enhancer activity measured by these assays are not subject to chromatin state, histone marks, and the native 3D-contacts, all of which may mask or enhance the measured enhancer activity of the DNA fragment *in vivo*[41,47]. To address this issue, we plan to generate homogenous cell lines containing specific variants of a set of TREs (i.e., a few dozen that show effects, either up- or down-regulation, on enhancer activity in **Aim 2**) using the CRISPR/Cas9 system in K562 cells[62]. Once verified, these mutant TRE cell lines will be compared to unmodified K562 cells. We will confirm the activity of selected variants of enhancers (from **Aim 2**) at native loci. In these engineered cells, we will perform PRO-seq to examine with high sensitivity the transcription at the variant TREs as well as all TREs and transcription units genome-wide. This will enable us to assess in an unbiased manner the role of elements of enhancers and promoters in the regulatory crosstalk between nearby and distant gene promoters and enhancers. To obtain an unbiased analysis of long-distance interactions of these enhancers, we will perform 4C experiments with the enhancer in question as the anchor site to measure broadly its interactions. Finally, we will test effects of these mutant enhancers on transcription factor binding and local histone marks at these genomic points of enhancer interaction by performing targeted ChIP-qPCR experiments. This approach rigorously examines our mutated TREs that show the most robust phenotypes in STROBE-seq, and thereby will define critical features of enhancers in executing their functions, and provide insights on how noncoding variants can impact gene regulation.

c.2.2.6. Potential pitfalls and alternative approaches. K562, though derived from a leukemia patient, is a lymphoblastoid Tier 1 cell line in the ENCODE project and it also has all necessary data available in the pattern of divergent transcription by GRO-cap[57], histone modifications[63], and DHS[64]. Moreover, >70% transfection efficiency can be achieved by electroporation in K562 cells, which makes it one of the best options for our proposed study. We have extensive experience with GRO-seq[57,65-75]; therefore, we are highly confident that our STROBE-seq experiments, an integration of STARR-seq and GRO-seq, will be successful. If we run into an unforeseen problem, we can always fall back on the standard STARR-seq protocol.

Furthermore, recent public patent applications suggest that Illumina is exploring long-read sequencing strategies related to the CHAIN-seq approach described here that may offer an alternative solution to studying large enhancer elements with STROBE-seq.

A confounding factor in STARR-seq and related applications, including STROBE-seq, is the interplay of the candidate enhancer, which is embedded in the transcription unit, and the core promoter used to drive the transcription of the transgene. Active enhancers are themselves transcribed, and the relative strength of transcription emanating from the enhancer (divergent transcription going away from the enhancer in both directions) and the promoter (towards the downstream enhancer fragment) may interfere with the enhancer activity. This interference may arise through modifications of histones that put elongation marks on site of transcription initiation[76]. In STARR-seq and STROBE-seq the enhancers are intragenic and therefore subject to promoter crosstalk, while native intergenic enhancers are immune from this crosstalk. High-throughput luciferase assays with an intergenic cloning site will be used to screen for such confounding effects, as will analysis at native sites of the most interesting mutation phenotypes in **c.2.2.5**.

c.3. Specific Aim 3. ReEnAct: a computational-experimental-integrated iterative learning framework to prioritize impactful non-coding variants

We plan to convert and extend the current FunSeq prototype from its focus on somatic variants to allow the identification of germline variants associated with large gene expression changes. Our new approach called ReEnAct (Regulation of Enhancer Activity) will iteratively create a model to predict high impact variants within regulatory regions of the genome (**Fig. 11**). It will have several features tailoring it to germline analysis, including 1) identifying functional sites among the conserved regions of the human genome and 2) investigating allelic elements. We will iteratively train and test our model on the results of the STROBE-seq experiments to refine its parameters.

c.3.1. Preliminary Studies

c.3.1.1. We have experience in annotating non-coding regions of the genome, including both TF-binding sites and non-coding RNAs. Our proposed work is based on our past experience in non-coding

genome annotation, as part of our 10-year history with the ENCODE and modENCODE projects. Our TF work includes the development of methods to define the binding peaks of TFs[77], prediction of a TF's target genes[78], and new machine learning techniques[79]. Furthermore, we developed methods that integrate ChIP-seq, chromatin, conservation, sequence and gene annotation data to identify gene-distal enhancers[80], which we have partially validated[81]. We also constructed linear and non-linear models that utilize TF binding and histone modification signals to accurately predict the transcriptional output of a gene in different cell types of several organisms including yeast, worm, fly, and human[22,82-85]. We have also constructed regulatory networks for human and model organisms[86,87], and completed many analyses on them[22,81,86,88-101].

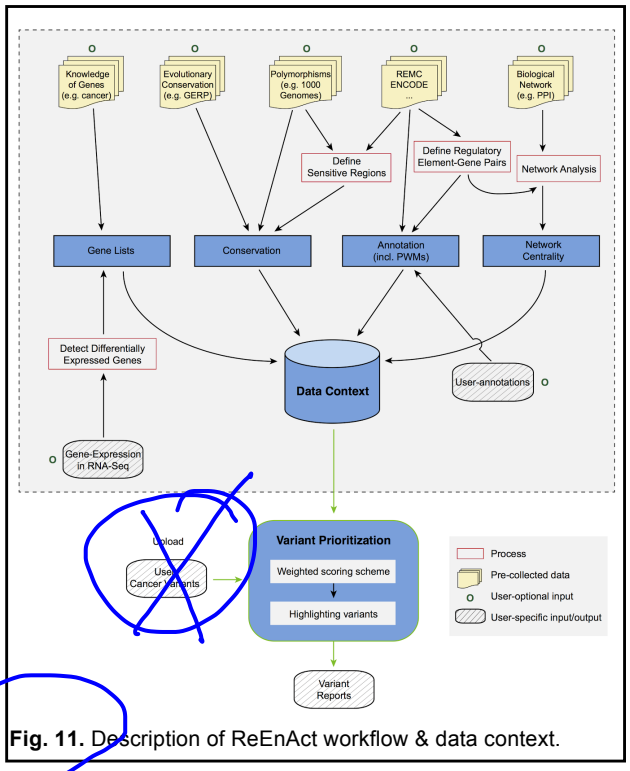

Fig. 11. Description of ReEnAct workflow & data context.

c.3.1.2. We have experience in allelic analyses. A specific class of regulatory variants is one that is related to allele-specific events. These are variants that are associated with allele-specific binding (ASB), particularly for transcription factors or DNA-binding proteins, and allele-specific expression (ASE)[102,103]. We have previously developed a tool, AlleleSeq,[99] for the detection of candidate variants associated with ASB and ASE. Using this we have generated comprehensive lists of allelic variants for ENCODE and 1000 Genomes and found that allelic variants are under differential selection from non-allelic ones[86,97,104]. By constructing regulatory networks based on ASB of TFs and ASE of their target genes, we further revealed substantial coordination between allele-specific binding and expression[86]. Furthermore, we have constructed a personal diploid genome and transcriptome of NA12878[105].

c.3.1.3. We have experience in relating annotation to variation: the FunSeq pipeline. We have extensively analyzed patterns of variation in non-coding regions, along with their coding targets[81,86,106]. In recent studies[97,107], we have integrated and extended these methods to develop a prioritization pipeline called FunSeq (Fig 2). It identifies sensitive and ultra-sensitive regions (i.e., those annotations under strong selective pressure, as determined using genomes from many individuals from diverse populations). FunSeq links each non-coding mutation to target genes, and prioritizes such variants based on scaled network connectivity. It identifies deleterious variants in many non-coding functional elements, including TF binding sites, enhancer elements, and regions of open chromatin corresponding to DNase I hypersensitive sites. Integrating large-scale data from various resources (including ENCODE and The 1000 Genomes Project) with cancer genomics data, our method is able to prioritize the known TERT promoter driver mutations. Using FunSeq, we identified ~100 non-coding candidate drivers in ~90 WGS medulloblastoma, breast and prostate cancer samples[97]. Drawing on this experience, we are currently co-leading the ICGC PCAWG-2 (analysis of mutations in regulatory regions) group.

c.3.2. Research Design

c.3.2.1. Prioritizing non-coding elements from polymorphism data. In order to identify variants that have a significant impact on gene regulation, we will use both intra-human variation data (from The 1000 Genomes Project) as well as cross-species evolutionary conservation (using classical measures such as GERP score[108]).

We will first update the TF binding non-coding elements from the original FunSeq approach. Due to the development of a number of massively parallel assays for identifying regulatory regions in the genomes, we

have been able to identify the epigenetic signatures underpinning active enhancers. We will use this information to make better enhancer predictions and utilize information provided by the Epigenome Roadmap [109-111], and more recently from ENCODE projects. In particular, we will develop a new machine-learning framework that combines pattern recognition within the signal of various epigenomic features and transcription of enhancer RNA (eRNA) with sequence-based features to predict active enhancers across different brain regions and other tissues in the Epigenome Roadmap project.

c.3.2.2. Identifying high-impact mutations: breaking & creating motifs. We will use motif breakers and formers to define loss-of- and gain-of-function events, respectively, as these events are more likely to have deleterious consequences[12,13,41,97,106,112,113]. Variants altering the position-weight matrix (PWM) scores for TF binding sites could potentially either decrease (loss-of-function) or increase (gain-of-function) the binding strength of TFs. A key improvement that we plan to utilize is to employ ancestral alleles to get a more accurate determination of these events.

c.3.2.3. Variant prioritization based on allelic activity. Allele-specific variants potentially provide the most direct readout of the functional impact of a variant. For example, if we can associate the differential binding effect of a particular transcription factor with different alleles of an SNV, then we can identify loci that have potential functional impacts in regulation. However, because allelic variants are enriched for rare variants[23], it will be difficult to match the specific variants in a personal genome of interest to prioritize against those earlier determined to be allelic in a functional genomics experiment on a cell line. Hence, instead of prioritizing by the direct overlap of allelic variants, we need to prioritize by the presence of allelic variants within 'allelic elements', or allelic regions in the genome (**Fig. 12**).

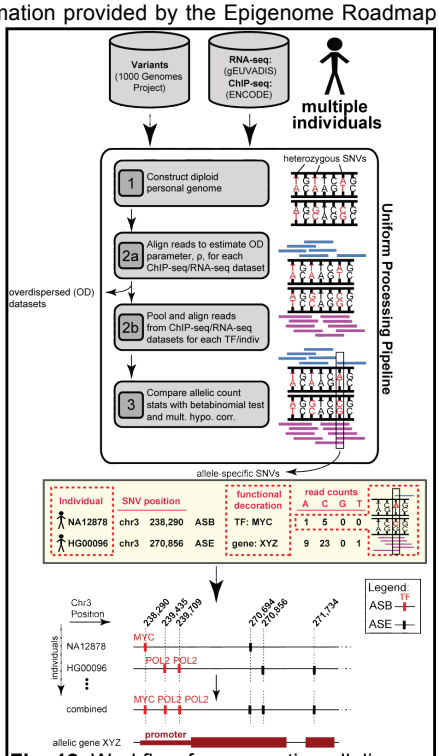

Fig. 12. Workflow for generating allelic variants and elements.

c.3.2.4. Identifying likely target genes for distal regulatory elements & assessing the impact of variants on network connectivity. To interpret the likely functional consequences of non-coding variants, we will comprehensively define associations between many non-coding regulatory elements and their target protein-coding genes. The correlation between enhancer and promoter activity across the ENCODE cell-lines and different tissues will be used to identify significant associations between regulatory elements and candidate target genes, as done by Yip et al[80]. A single regulatory variant may affect the expression of multiple genes, either because it directly regulates multiple genes or because the target gene is itself a regulatory factor.

c.3.2.5. We will use a unified weighted scoring scheme for combining all ReEnAct features to prioritize variants. To integrate the various features mentioned above, we plan to elaborate the weighting system in FunSeq[107]. Constrained by selective pressure, common variations tend to arise in functionally unimportant regions. Thus, features that are enriched with common polymorphisms are less likely to contribute to the deleteriousness of variants and are weighted less. In general, features can be classified into two classes: discrete (e.g., within or outside of a given functional annotation) and continuous (e.g., the PWM change in 'motif-breaking'). We will weigh these two sets of features with different strategies.

For each discrete feature d, we calculate the probability p_d that it overlaps with common polymorphisms. We then calculate the information content to denote the value of discrete features $s_d = 1 + p_d * log_2 p_d + (1 - p_d) * log_2(1 - p_d)$.

The situation is more complex for continuous features, as different feature values have different probabilities of being observed in natural polymorphisms. Thus, one weight cannot suffice for varied feature values. For a continuous feature c, which is associated with a value v_c, the probability $p_c^{v_c}$ is firstly estimated using common variants: $p_c^{v_c} = \frac{\#common\ variant\ v \geq v_c}{\#common\ variant}$. The score of continuous feature is defined as $s_c^{v_c} = 1 + p_c^{v_c} * log_2 p_c^{v_c} + (1 - p_c^{v_c}) * log_2(1 - p_c^{v_c})$.

Haiyuan Yu 1/9/2016 1:30 PM
Deleted: .

Haiyuan Yu 1/9/2016 1:29 PM
Deleted:

The ReEnAct score ($\boldsymbol{RS}$) is calculated as $\boldsymbol{RS} = \sum_d \theta_d s_d + \sum_c \theta_c s_c^{v_c} = \langle \boldsymbol{\theta}, S \rangle$. We will also incorporate the feature dependency structure when calculating the scores by removing redundant features using feature selection or by performing dimensionality reduction.

c.3.2.6. Three-stage Computational-Experimental-integrated Iterative learning with real-time experimental parameter optimization. Let $\boldsymbol{\theta^{(0)}} = (\theta_1^{(0)}, \theta_2^{(0)}, \ldots, \theta_m^{(0)})$ represent the initial feature parameters chosen randomly, where m is the number of features. $\boldsymbol{\theta}$ will be optimized using an iterative learning scheme by incorporating new experimental information produced in **Aims 1** and **2**. Because of the high throughput of MegaMut and STROBE-seq, our strategy is to implement for the first time a three-stage iterative learning scheme: the first stage initial learning, the second stage real-time experimental parameter optimization, and the third stage final assessment (**Fig. 1**).

In the first stage, we will randomly select ~2,000 active enhancers in K562 cells as defined by histone modifications[63], DNase I Hypersensitivity mapping[64], and GRO-cap signal[57]. We will first generate the WT clones of these enhancers using Clone-seq. Then, we will select 5 random variants on each enhancers and generate all ~10,000 variant clones through MegaMut. Their effects on enhancer activities will be quantified by STROBE-seq. Starting from the initial tuned $\boldsymbol{\theta^{(0)}}$, we tune $\boldsymbol{\theta^{(1)}}$ according to STROBE-seq results of ~10,000 variants in the first stage. For a specific variant v, we define y_v as Bernoulli distributed random variable with $y_v = 1$ indicates that v is functional. Then expectation of y_v can be predicted through a logistic regression. Specifically, $\text{logit}(P(y_v = 1)) = -k * (\boldsymbol{RS}_v - a) = -k * \left(\sum_m \theta_m * s_{v,m} - a\right)$ (k, a are scaling parameters). To update $\boldsymbol{\theta^{(0)}}$ with experimental validation results $\boldsymbol{Y}$, we implement Bayes' rule: $P(\boldsymbol{\theta}|\boldsymbol{Y}) \propto P(\boldsymbol{Y}|\boldsymbol{\theta})P(\boldsymbol{\theta})$. We will use MCMC (Monte Carlo Markov Chain) sampling to search over the parameter space and find the most probable $\boldsymbol{\theta^{(1)}}$. We will predict the functional impact of all variants, $P(y_v = 1)$, using $\boldsymbol{\theta^{(1)}}$.

To select the variants for the first round of experimental optimization in the second stage, we will clone 200 untested enhancers and choose 1,000 variants (400 with predicted high impact, 200 with medium impact, and 400 with low impact). We will use MCMC to find the most probable $\boldsymbol{\theta^{(2\_1)}}$ and update all the $P(y_v = 1)$ scores. In subsequent rounds of the second stage, we will choose the top 200 enhancers (untested enhancers will be cloned through Clone-seq first) that contain variants v with the largest change in absolute logarithm odds ratio (AOR): $|\log(\text{OR})| = \left|\text{logit}\left(\text{P}\left(y_v^{(1)} = 1|\boldsymbol{\theta^{(1)}}\right)\right) - \text{logit}\left(\text{P}\left(y_v^{(0)} = 1|\boldsymbol{\theta^{(0)}}\right)\right)\right| = \left|-k\left(\boldsymbol{RS}^{(1)} - \boldsymbol{RS}^{(0)}\right)\right|$. We will clone and test 1,000 top AOR variants in these enhancers, and perform parameter optimization through MCMC. After the 6th round, we will obtain the final parameter $\boldsymbol{\theta^{(2)}}$ for ReEnAct. Depending on the results, we may add more rounds to this stage, or continue to the third stage.

In the third stage of final assessment, we will select 1,000 variants (400 with predicted high impact, 200 with medium impact, and 400 with low impact) on previously cloned enhancers and 4,000 variants (1,500 high impact, 1,000 medium impact, and 1,500 low impact) on 800 untested enhancers. We will measure their impact on enhancer activities quantitatively through STROBE-seq, which will be used to comprehensively evaluate the performance of ReEnAct. Overall, we will generate >3,000 WT enhancer clones and examine >20,000 noncoding variants using MegaMut + STROBE-seq.

c.3.2.7. Potential pitfalls and alternative approaches. The "three-stage" iterative learning structure provides a way to iteratively include experimental information into the machine-learning framework. However, since there are several rounds of iterative refinement, the final model may become overfit to the training data. To identify possible overfitting, we will use a learning curve (training error and validation error as functions of the number of training points). Since the iterative refinement is performed only on the training data, a traditional learning curve cannot be drawn. However, this can be circumvented by performing an internal cross-validation on each of the rounds of iterative refinement: we randomly assign a subset of the 1,000 variants tested in any round as training and the rest as validation just to compute the learning curve. If the training error rate (computed only on the subset of the 1,000 variants) is significantly lower than the validation error rate (computed using the rest of the 1,000 variants) even after including a large number of training points, the model is likely overfitted. If that is the case, we will attempt to avoid overfitting by a) including only a random subset of points per run; b) using a L1 regularization term to decrease the variance[118].

It is also possible that our model will be underfit after 6 rounds in the second stage. Our initial proposal is to test 1,000 variants per run over 6 runs in Year 3. However, if the learning curve shows that our model is underfitted (both training and validation error rate are high), we can fit a better model by a) using a L2 regularization term and/or reducing the penalty of the regularization term to decrease the bias; b) doing more runs in the second stage if necessary[118].

Haiyuan Yu 1/9/2016 3:58 PM
Deleted: ***c.3.2.6. Initial tuning based on publicly available datasets.*** To perform the initial round of performance assessment and parameter tuning, we plan to use publicly available datasets from various resources. The Human Gene Mutation Database (HGMD)[114] and ClinVar[115] catalogue large numbers of regulatory disease-causing mutations. Several high-throughput technologies have also been developed to test the phenotypic impacts of non-coding genomic variants. For example, Kwasnieski et al used CRE-seq[116] to assay over 1,000 single- and double-nucleotide mutations in promoter regions. Kheradpour et al.[41] used MPRA to test variants affecting regulatory motifs in over 2,000 human enhancers. We will utilize these datasets to perf ... [1]

Haiyuan Yu 1/9/2016 3:02 PM
Deleted: …parameters chosen ran ... [2]

Haiyuan Yu 1/9/2016 3:04 PM
Formatted: Highlight

Haiyuan Yu 1/9/2016 3:04 PM
Deleted: feature

Haiyuan Yu 1/9/2016 3:58 PM
Deleted: as part of this grant

Haiyuan Yu 1/9/2016 3:58 PM
Formatted: ... [3]

Haiyuan Yu 1/9/2016 3:59 PM
Deleted: O…r strategy is to implem ... [4]

Haiyuan Yu 1/9/2016 4:05 PM
Formatted: Highlight

Haiyuan Yu 1/9/2016 4:05 PM
Deleted: Given we $N(N > 20{,}000)$ ... [5]

Haiyuan Yu 1/9/2016 3:07 PM
Deleted: >

Haiyuan Yu 1/9/2016 4:24 PM
Deleted: experimental validation in ... [6]

Haiyuan Yu 1/9/2016 3:12 PM
Formatted: Font:Italic, Underline

Haiyuan Yu 1/9/2016 4:36 PM
Deleted: for the second stage of exp ... [7]

Haiyuan Yu 1/9/2016 4:40 PM
Formatted: Superscript

Haiyuan Yu 1/9/2016 4:40 PM
Deleted: the second stage ... [8]

Haiyuan Yu 1/9/2016 4:39 PM
Deleted: :…final assessment, we w ... [9]

Haiyuan Yu 1/9/2016 3:04 PM
Deleted: ***8.… Potential pitfalls a*** ... [10]

Haiyuan Yu 1/9/2016 6:10 PM
Deleted: the validation error stead ... [11]

Haiyuan Yu 1/9/2016 4:50 PM
Formatted ... [12]

Haiyuan Yu 1/9/2016 4:50 PM
Deleted: [SKL: L1, L2 is used to ... [13]

REFERENCES CITED

1. Cruchaga, C.*, et al.* Rare coding variants in the phospholipase D3 gene confer risk for Alzheimer's disease. *Nature* **505**, 550-554 (2014).
2. MacArthur, D.G.*, et al.* A systematic survey of loss-of-function variants in human protein-coding genes. *Science* **335**, 823-828 (2012).
3. Cox, A.*, et al.* A common coding variant in CASP8 is associated with breast cancer risk. *Nat Genet* **39**, 352-358 (2007).
4. Momozawa, Y.*, et al.* Resequencing of positional candidates identifies low frequency IL23R coding variants protecting against inflammatory bowel disease. *Nat Genet* **43**, 43-47 (2011).
5. Kimchi-Sarfaty, C.*, et al.* A "silent" polymorphism in the MDR1 gene changes substrate specificity. *Science* **315**, 525-528 (2007).
6. Musunuru, K.*, et al.* From noncoding variant to phenotype via SORT1 at the 1p13 cholesterol locus. *Nature* **466**, 714-719 (2010).
7. Ward, L.D. & Kellis, M. Interpreting noncoding genetic variation in complex traits and human disease. *Nat Biotechnol* **30**, 1095-1106 (2012).
8. De Gobbi, M.*, et al.* A regulatory SNP causes a human genetic disease by creating a new transcriptional promoter. *Science* **312**, 1215-1217 (2006).
9. Maller, J.*, et al.* Common variation in three genes, including a noncoding variant in CFH, strongly influences risk of age-related macular degeneration. *Nat Genet* **38**, 1055-1059 (2006).
10. Cheung, V.G. & Spielman, R.S. Genetics of human gene expression: mapping DNA variants that influence gene expression. *Nat Rev Genet* **10**, 595-604 (2009).
11. Li, Q.*, et al.* Integrative eQTL-based analyses reveal the biology of breast cancer risk loci. *Cell* **152**, 633-641 (2013).
12. Huang, F.W.*, et al.* Highly recurrent TERT promoter mutations in human melanoma. *Science* **339**, 957-959 (2013).
13. Horn, S.*, et al.* TERT promoter mutations in familial and sporadic melanoma. *Science* **339**, 959-961 (2013).
14. Tournamille, C., Colin, Y., Cartron, J.P. & Le Van Kim, C. Disruption of a GATA motif in the Duffy gene promoter abolishes erythroid gene expression in Duffy-negative individuals. *Nat Genet* **10**, 224-228 (1995).
15. McCarroll, S.A.*, et al.* Deletion polymorphism upstream of IRGM associated with altered IRGM expression and Crohn's disease. *Nat Genet* **40**, 1107-1112 (2008).
16. Verlaan, D.J.*, et al.* Targeted screening of cis-regulatory variation in human haplotypes. *Genome Res* **19**, 118-127 (2009).
17. Ponting, C.P. & Lunter, G. Signatures of adaptive evolution within human non-coding sequence. *Hum Mol Genet* **15 Spec No 2**, R170-175 (2006).
18. Ghildiyal, M. & Zamore, P.D. Small silencing RNAs: an expanding universe. *Nat Rev Genet* **10**, 94-108 (2009).
19. Kleinjan, D.A. & van Heyningen, V. Long-range control of gene expression: emerging mechanisms and disruption in disease. *Am J Hum Genet* **76**, 8-32 (2005).
20. ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57-74 (2012).
21. modENCODE Consortium*, et al.* Identification of functional elements and regulatory circuits by Drosophila modENCODE. *Science* **330**, 1787-1797 (2010).
22. Gerstein, M.B.*, et al.* Integrative analysis of the Caenorhabditis elegans genome by the modENCODE project. *Science* **330**, 1775-1787 (2010).
23. Lappalainen, T.*, et al.* Transcriptome and genome sequencing uncovers functional variation in humans. *Nature* **501**, 506-511 (2013).
24. Montgomery, S.B.*, et al.* Transcriptome genetics using second generation sequencing in a Caucasian population. *Nature* **464**, 773-777 (2010).

25. Battle, A.*, et al.* Characterizing the genetic basis of transcriptome diversity through RNA-sequencing of 922 individuals. *Genome Res* **24**, 14-24 (2014).
26. Ernst, J. & Kellis, M. Large-scale imputation of epigenomic datasets for systematic annotation of diverse human tissues. *Nat Biotechnol* **33**, 364-376 (2015).
27. Baran, Y.*, et al.* The landscape of genomic imprinting across diverse adult human tissues. *Genome Res* (2015).
28. Won, K.-J.*, et al.* Comparative annotation of functional regions in the human genome using epigenomic data. *Nucleic Acids Res* **41**, 4423-4432 (2013).
29. Mackay, T.F., Stone, E.A. & Ayroles, J.F. The genetics of quantitative traits: challenges and prospects. *Nat Rev Genet* **10**, 565-577 (2009).
30. Carlborg, O. & Haley, C.S. Epistasis: too often neglected in complex trait studies? *Nat Rev Genet* **5**, 618-625 (2004).
31. Lappalainen, T., Montgomery, S.B., Nica, A.C. & Dermitzakis, E.T. Epistatic selection between coding and regulatory variation in human evolution and disease. *Am J Hum Genet* **89**, 459-463 (2011).
32. Marchini, J., Donnelly, P. & Cardon, L.R. Genome-wide strategies for detecting multiple loci that influence complex diseases. *Nat Genet* **37**, 413-417 (2005).
33. Haraksingh, R.R. & Snyder, M.P. Impacts of variation in the human genome on gene regulation. *J Mol Biol* **425**, 3970-3977 (2013).
34. Hemani, G.*, et al.* Detection and replication of epistasis influencing transcription in humans. *Nature* **508**, 249-253 (2014).
35. Sanseau, P.*, et al.* Use of genome-wide association studies for drug repositioning. *Nat Biotechnol* **30**, 317-320 (2012).
36. Cao, C. & Moult, J. GWAS and drug targets. *BMC Genomics* **15 Suppl 4**, S5 (2014).
37. Weinshilboum, R. & Wang, L. Pharmacogenomics: bench to bedside. *Nat Rev Drug Discov* **3**, 739-748 (2004).
38. Xie, L.*, et al.* Towards structural systems pharmacology to study complex diseases and personalized medicine. *PLoS Comput Biol* **10**, e1003554 (2014).
39. Collins, F.S. & Varmus, H. A new initiative on precision medicine. *N Engl J Med* **372**, 793-795 (2015).
40. Patwardhan, R.P.*, et al.* Massively parallel functional dissection of mammalian enhancers in vivo. *Nat Biotechnol* **30**, 265-270 (2012).
41. 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).
42. 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).
43. Fowler, D.M.*, et al.* High-resolution mapping of protein sequence-function relationships. *Nature methods* **7**, 741-746 (2010).
44. 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).
45. Hiatt, J.B., Patwardhan, R.P., Turner, E.H., Lee, C. & Shendure, J. Parallel, tag-directed assembly of locally derived short sequence reads. *Nat Methods* **7**, 119-122 (2010).
46. 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).
47. Arnold, C.D.*, et al.* Genome-wide quantitative enhancer activity maps identified by STARR-seq. *Science* **339**, 1074-1077 (2013).
48. Whyte, W.A.*, et al.* Master transcription factors and mediator establish super-enhancers at key cell identity genes. *Cell* **153**, 307-319 (2013).
49. Hnisz, D.*, et al.* Super-enhancers in the control of cell identity and disease. *Cell* **155**, 934-947 (2013).
50. Pickrell, J.K.*, et al.* Understanding mechanisms underlying human gene expression variation with RNA sequencing. *Nature* **464**, 768-772 (2010).
51. Montgomery, S.B., Lappalainen, T., Gutierrez-Arcelus, M. & Dermitzakis, E.T. Rare and common regulatory variation in population-scale sequenced human genomes. *PLoS Genet* **7**, e1002144 (2011).

52. 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).
53. 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).
54. Firnberg, E. & Ostermeier, M. PFunkel: efficient, expansive, user-defined mutagenesis. *PLoS One* **7**, e52031 (2012).
55. Kowalsky, C.A., *et al.* High-resolution sequence-function mapping of full-length proteins. *PLoS One* **10**, e0118193 (2015).
56. Kunkel, T.A. Rapid and Efficient Site-Specific Mutagenesis without Phenotypic Selection. *Proceedings of the National Academy of Sciences of the United States of America* **82**, 488-492 (1985).
57. 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).
58. Zabidi, M.A., *et al.* Enhancer-core-promoter specificity separates developmental and housekeeping gene regulation. *Nature* **518**, 556-559 (2015).
59. Picelli, S., *et al.* Tn5 transposase and tagmentation procedures for massively scaled sequencing projects. *Genome Res* **24**, 2033-2040 (2014).
60. Kwak, H., Fuda, N.J., Core, L.J. & Lis, J.T. Precise Maps of RNA Polymerase Reveal How Promoters Direct Initiation and Pausing. *Science* **339**, 950-953 (2013).
61. Treisman, R. & Maniatis, T. Simian virus 40 enhancer increases number of RNA polymerase II molecules on linked DNA. *Nature* **315**, 73-75 (1985).
62. Greenfeld, H., *et al.* TRAF1 Coordinates Polyubiquitin Signaling to Enhance Epstein-Barr Virus LMP1-Mediated Growth and Survival Pathway Activation. *PLoS pathogens* **11**, e1004890 (2015).
63. Roadmap Epigenomics Consortium, *et al.* Integrative analysis of 111 reference human epigenomes. *Nature* **518**, 317-330 (2015).
64. Neph, S., *et al.* An expansive human regulatory lexicon encoded in transcription factor footprints. *Nature* **489**, 83-90 (2012).
65. 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).
66. Danko, C.G., *et al.* Identification of active transcriptional regulatory elements from GRO-seq data. *Nat Methods* **12**, 433-438 (2015).
67. 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).
68. 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).
69. 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).
70. 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).
71. 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).
72. Hah, N., *et al.* A rapid, extensive, and transient transcriptional response to estrogen signaling in breast cancer cells. *Cell* **145**, 622-634 (2011).
73. Min, I.M., *et al.* Regulating RNA polymerase pausing and transcription elongation in embryonic stem cells. *Genes Dev* **25**, 742-754 (2011).
74. Larschan, E., *et al.* X chromosome dosage compensation via enhanced transcriptional elongation in Drosophila. *Nature* **471**, 115-118 (2011).
75. 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).
76. Berger, S.L. The complex language of chromatin regulation during transcription. *Nature* **447**, 407-412 (2007).

77. Rozowsky, J.*, et al.* PeakSeq enables systematic scoring of ChIP-seq experiments relative to controls. *Nat Biotechnol* **27**, 66-75 (2009).
78. Cheng, C., Min, R. & Gerstein, M. TIP: a probabilistic method for identifying transcription factor target genes from ChIP-seq binding profiles. *Bioinformatics* **27**, 3221-3227 (2011).
79. Yip, K.Y. & Gerstein, M. Training set expansion: an approach to improving the reconstruction of biological networks from limited and uneven reliable interactions. *Bioinformatics* **25**, 243-250 (2009).
80. Yip, K.Y., Alexander, R.P., Yan, K.-K. & Gerstein, M. Improved reconstruction of in silico gene regulatory networks by integrating knockout and perturbation data. *PLoS One* **5**, e8121 (2010).
81. Yip, K.Y.*, et al.* Classification of human genomic regions based on experimentally determined binding sites of more than 100 transcription-related factors. *Genome Biol* **13**, R48 (2012).
82. Cheng, C., Shou, C., Yip, K.Y. & Gerstein, M.B. Genome-wide analysis of chromatin features identifies histone modification sensitive and insensitive yeast transcription factors. *Genome Biol* **12**, R111 (2011).
83. Cheng, C.*, et al.* A statistical framework for modeling gene expression using chromatin features and application to modENCODE datasets. *Genome Biol* **12**, R15 (2011).
84. Cheng, C. & Gerstein, M. Modeling the relative relationship of transcription factor binding and histone modifications to gene expression levels in mouse embryonic stem cells. *Nucleic Acids Res* **40**, 553-568 (2012).
85. Cheng, C.*, et al.* Understanding transcriptional regulation by integrative analysis of transcription factor binding data. *Genome Res* **22**, 1658-1667 (2012).
86. Gerstein, M.B.*, et al.* Architecture of the human regulatory network derived from ENCODE data. *Nature* **489**, 91-100 (2012).
87. Nègre, N.*, et al.* A cis-regulatory map of the Drosophila genome. *Nature* **471**, 527-531 (2011).
88. Cheng, C.*, et al.* Construction and analysis of an integrated regulatory network derived from high-throughput sequencing data. *PLoS Comput Biol* **7**, e1002190 (2011).
89. Yan, K.-K., Fang, G., Bhardwaj, N., Alexander, R.P. & Gerstein, M. Comparing genomes to computer operating systems in terms of the topology and evolution of their regulatory control networks. *Proc Natl Acad Sci U S A* **107**, 9186-9191 (2010).
90. Yu, H., Greenbaum, D., Xin Lu, H., Zhu, X. & Gerstein, M. Genomic analysis of essentiality within protein networks. *Trends Genet* **20**, 227-231 (2004).
91. Yu, H., Zhu, X., Greenbaum, D., Karro, J. & Gerstein, M. TopNet: a tool for comparing biological sub-networks, correlating protein properties with topological statistics. *Nucleic Acids Res* **32**, 328-337 (2004).
92. Yu, H., Kim, P.M., Sprecher, E., Trifonov, V. & Gerstein, M. The importance of bottlenecks in protein networks: correlation with gene essentiality and expression dynamics. *PLoS Comput Biol* **3**, e59 (2007).
93. Luscombe, N.M.*, et al.* Genomic analysis of regulatory network dynamics reveals large topological changes. *Nature* **431**, 308-312 (2004).
94. Gianoulis, T.A.*, et al.* Quantifying environmental adaptation of metabolic pathways in metagenomics. *Proc Natl Acad Sci U S A* **106**, 1374-1379 (2009).
95. Yu, H., Paccanaro, A., Trifonov, V. & Gerstein, M. Predicting interactions in protein networks by completing defective cliques. *Bioinformatics* **22**, 823-829 (2006).
96. Kim, P.M., Korbel, J.O. & Gerstein, M.B. Positive selection at the protein network periphery: evaluation in terms of structural constraints and cellular context. *Proc Natl Acad Sci U S A* **104**, 20274-20279 (2007).
97. Khurana, E.*, et al.* Integrative annotation of variants from 1092 humans: application to cancer genomics. *Science* **342**, 1235587 (2013).
98. Khurana, E., Fu, Y., Chen, J. & Gerstein, M. Interpretation of genomic variants using a unified biological network approach. *PLoS Comput Biol* **9**, e1002886 (2013).
99. Rozowsky, J.*, et al.* AlleleSeq: analysis of allele-specific expression and binding in a network framework. *Mol Syst Biol* **7**, 522 (2011).
100. Lörcher, U., Peters, J. & Kollath, J. [Changes in the lungs and pleura following chemoembolization of liver tumors with mitomycin-lipiodol]. *Rofo* **152**, 569-573 (1990).

101. Shou, C., *et al.* Measuring the evolutionary rewiring of biological networks. *PLoS Comput Biol* **7**, e1001050 (2011).
102. Pastinen, T. Genome-wide allele-specific analysis: insights into regulatory variation. *Nat Rev Genet* **11**, 533-538 (2010).
103. Birney, E., Lieb, J.D., Furey, T.S., Crawford, G.E. & Iyer, V.R. Allele-specific and heritable chromatin signatures in humans. *Hum Mol Genet* **19**, R204-209 (2010).
104. Djebali, S., *et al.* Landscape of transcription in human cells. *Nature* **489**, 101-108 (2012).
105. http://alleleseq.gersteinlab.org. Last accessed on 21st May 2015.
106. Mu, X.J., Lu, Z.J., Kong, Y., Lam, H.Y.K. & Gerstein, M.B. Analysis of genomic variation in non-coding elements using population-scale sequencing data from the 1000 Genomes Project. *Nucleic Acids Res* **39**, 7058-7076 (2011).
107. Fu, Y., *et al.* FunSeq2: a framework for prioritizing noncoding regulatory variants in cancer. *Genome Biol* **15**, 480 (2014).
108. Cooper, G.M., *et al.* Distribution and intensity of constraint in mammalian genomic sequence. *Genome Res* **15**, 901-913 (2005).
109. Roadmap Epigenomics Consortium, *et al.* Integrative analysis of 111 reference human epigenomes. *Nature* **518**, 317-330 (2015).
110. Ziller, M.J., *et al.* Dissecting neural differentiation regulatory networks through epigenetic footprinting. *Nature* **518**, 355-359 (2015).
111. Leung, D., *et al.* Integrative analysis of haplotype-resolved epigenomes across human tissues. *Nature* **518**, 350-354 (2015).
112. Killela, P.J., *et al.* TERT promoter mutations occur frequently in gliomas and a subset of tumors derived from cells with low rates of self-renewal. *Proc Natl Acad Sci U S A* **110**, 6021-6026 (2013).
113. Vinagre, J., *et al.* Frequency of TERT promoter mutations in human cancers. *Nat Commun* **4**, 2185 (2013).
114. Stenson, P.D., *et al.* Human Gene Mutation Database (HGMD): 2003 update. *Hum Mutat* **21**, 577-581 (2003).
115. Landrum, M.J., *et al.* ClinVar: public archive of relationships among sequence variation and human phenotype. *Nucleic Acids Res* **42**, D980-985 (2014).
116. Kwasnieski, J.C., Mogno, I., Myers, C.A., Corbo, J.C. & Cohen, B.A. Complex effects of nucleotide variants in a mammalian cis-regulatory element. *Proc Natl Acad Sci U S A* **109**, 19498-19503 (2012).
117. Kircher, M., *et al.* A general framework for estimating the relative pathogenicity of human genetic variants. *Nat Genet* **46**, 310-315 (2014).
118. Bunea, F. Honest variable selection in linear and logistic regression models via l1 and l1 + l2 penalization. *Electromic Journal of Statistics* **2**, 1153-1194 (2008).