Introduction (response to criticism of the previous submission)

This is a revision of a proposal that we originally submitted about a year and a half ago. The proposal was fairly well reviewed, with the committee expressing "high enthusiasm" for the "likelihood that a useful computational tool for prioritization of non-coding variants will come from the project" and "good potential for being applicable to a variety of traits and diseases." In the resubmission we attempt to address some critical feedback that our original proposal received, particularly that the "laboratory correlative aspects of [our] proposal are not as well developed as the detailed sophisticated computational aspects." In the revision we attempt to address criticism in a number of ways:

(a) We have split the original Aim 3, which was the experimental validation, into two aims, one focusing on high-throughput experimental characterization (new Aim 3) and the other on detailed validation (new Aim 4). Furthermore, in our new Aim 2, we describe how our variant prioritization pipeline can be iteratively tuned using high-throughput experimental characterization (new Aim 3) and then assessed with the detailed validation (new Aim 4).

(b) To help with our elaborated experimental plan, we have also added a new co-investigator, Dr. Haiyuan Yu, who was part of the original "1000 Genomes FIG/FunSeq collaboration" between Yale and Cornell, which developed the prototype of our approach described in the preliminary results. His lab brings significant expertise in terms of high-throughput validation. Dr. Yu will implement his newly-established massively-parallel site-directed mutagenesis pipeline, Clone-seq (see new Aim 3), to generate ~1200 specific mutations in ~600 regulatory elements in three rounds to tune the parameters of the our variant prioritization pipeline and comprehensively evaluate the performance of the computational pipeline. He regularly meets with Dr Rubin and Dr Gerstein and has a strong co-publication record with both.

(c) Note that the splitting of our original Aim 3 and the creation of the tuning approach allows us to deal with the different scales on which one can approach functional characterization of the genome. In particular, in new Aim 2, we deal with millions of computationally predicted functional sites; in new Aim 3, we perform high-throughput experimental characterization using Clone-seq and reporter assays and computational tuning on thousands of potential functional sites; and in new Aim 4, we carry out detailed experimental validation with CRISPR/Cas9 and other assays on individual variants. We believe this multi-scale approach, which we diagram in a new schematic (Fig 5), is an optimal way to harness both breadth and depth.

The experimental plan was also specifically criticized for not detailing how "cell context can be important" for validation assays and pipeline tuning. Here we address this issue by more explicitly focusing our validations on the prostate cancer system, using a prostate cancer cohort for the detailed validation and the LNCaP and PC3 prostate cancer cell lines for the high-throughput experiments. This will enable us to show how our variant prioritization scheme can be refined and evaluated for a specific disease context.

We were also criticized in the original submission regarding the large "computational load for handling all germline variants from TCGA and ICGC whole genome sequences." We addressed this by collaborating with the PCAWG germline variant calling group, enabling us to cut almost all in-house variant calling from the proposal. (See letter of collaboration form J Korbel, head of PCAWG-8, Pan Cancer Analysis Working Group #8.) We will also take advantage of recently published studies, in which individual groups called germline variants for large portions of the available TCGA and ICGC whole genome data (see Aim 2).

Feedback on our original proposal noted our "highly innovative computational [approach] for interpreting sequence variants" and described our "focus on rare variants [as] a strength as their role has often proven difficult to interpret." We have preserved these strengths of the previous submission. However, the time since our previous submission has allowed us to progress in our work and to formulate refined approaches for analysis. In particular, we have published the FunSeq2 paper[74], which focuses on somatic variants but which incorporates some of the ideas proposed in the original grant application. We also developed the LARVA to look for recurrent mutations and highly burdened regulatory elements, taking into account various functional genomics features such as replication timing. We have written this up as a publication for *Nucleic Acids Research*, now in revision. We have moved these originally proposed ideas into preliminary results and have updated our approach with new ideas focused on allelic activity, a consistent scoring framework for variants in all non-coding regions of the genome, and a new method for tuning our variant scoring scheme to experimental data.

A. Specific aims

Recent progress made by The ENCODE Consortium and The Epigenome Roadmap Project has provided detailed annotation of non-coding regions of the human genome, and whole-genome sequencing has identified large volumes of rare variants in such regions, thereby making this an opportune time to study rare variants in non-coding regions. Despite these opportunities, little effort has been invested in leveraging these resources to tackle problems in cancer-risk variant prioritization. Here, we detail our strategies to use patterns of natural polymorphism to prioritize the most impactful non-coding variants. In order to refine our approach, we will develop a tunable parameterization scheme, in conjunction with iterative experimentation.

Specifically, we will adapt our FunSeq pipeline for scoring cancer somatic variants to create a tool for prioritizing rare germline variants, eleVAR (**ele**vating germline **VAR**iants). This tool will have the expanded capability to score variants in non-coding DNA and RNA regions uniformly, and it will contain the flexible weighting scheme that will be subject to successive rounds of validation and tuning. We will first do this in a generic fashion using publicly available data, followed by our own validation experiments in prostate cancer cell lines and patient samples to tune and assess a targeted version of eleVAR, as a prototype for how our approach can be focused on particular disease.

Aim 1. Adapt our existing tool for prioritizing somatic variants (FunSeq) to create a generalizable approach for prioritizing impactful non-coding variants (eleVAR). 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 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 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.

Aim 2. Implement eleVAR pipeline and develop a workflow for tuning and assessing performance, focusing on prostate cancer as a test case for a specific disease. In particular, **(a)** we will implement eleVAR as an efficient software package, with separate modules for building the data context from annotations, parameter tuning, and scoring variants. **(b)** We will use eleVAR to generate a generic list of prioritized variants from the PCAWG germline variants. **(c)** We will develop a Bayesian update approach to tune the eleVAR parameters in response to validation data. We will then carry out tuning, **(d)** first based on publicly available validation data, such as known mutations causing disease phenotypes or functional changes in enhancer activity assays. Then we will tune based on our own high-throughput luciferase reporter experiments in prostate cancer cells (see Aim 3). **(e)** We will make our code and tuning results publically available on a project website.

Aim 3. High-throughput experimental characterization of ~1200 variants using Clone-seq and luciferase reporter assays. 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 250 genomic elements, and use our newly-developed massively-parallel Clone-seq pipeline to generate two variants, one highly prioritized (predicted to be deleterious by eleVAR) and one with a lower score (500 total variants/round). We will assess impact of variants on gene regulation using high-throughput luciferase reporter assays, comparing wild type and predicted high and low impact variants. **(b)** In the final round, purely for assessment, we will choose another 100 genomic elements and two variants with high and low scores on each element and generate clones using Clone-seq. In total, we generate clones for ~600 WT genomic elements and ~1200 variants, which will allow for a comprehensive evaluation of the eleVAR performance.

Aim 4. Detailed experimental validation of a few non-coding variants from eleVAR. We will perform a detailed in-depth experimental validation on 6 representative positive variants from eleVAR after tuning and high-throughput validation. **(a)** We will use TaqMan assays to genotype these variants in ~4,000 samples from a cohort of prostate cancer patients. **(b)** We will further evaluate them for biochemical validation by introducing them into their endogenous loci using the CRISPR/Cas-9 system. We will then assay their downstream effects on gene-expression using real-time quantitative PCR, as well as cell viability, migratory potential (for metastasis), and transcription factor binding (ChIP and EMSA).

Resource Sharing Plan

A comprehensive list of the rare germline variants which are prioritized by our eleVAR pipeline will be made available. Data on variants from 2000 genomes will be submitted to dbGaP in VCF format. Along with the variants themselves, we will provide the important accessory data from the eleVAR ranking. This will include whether the variants lie in sensitive regions, regulatory network hubs, transcription-factor binding sites, and other elements likely to impact phenotype. Also we will notate whether a given variant is a gain- or loss-of-function variant at binding sites. For variants in enhancers or binding sites, we will provide data on their associated gene(s), such as their genomic loci and ensembl gene IDs. We will also provide metrics from the weighting scheme for each variant.

Among the rare variants that we prioritize and to which we assign features, we will publicly release about a thousand in flat text files on our eleVAR website (see below). This amounts to less than one variant per genome. As a result, we emphasize that this will not constitute any real concerns regarding privacy -- though we will modify this if the NIH feels otherwise.

Among these variants, we will experimentally assay about 100, as stated in our proposal. The results gained from the experimental assays for validating variants (such as the TaqMan assays) will be released in the form of flat text files. We plan on releasing this data to the public once the work analyzing the data is published, or three months after the award period ends (whichever is shorter). Again, given that this is substantially less that one variant per genome, this should not constitute a problem as far as privacy is concerned.

We will create a project website (elevar.gersteinlab.org) that will house our source code and our validation data. As there is no large-scale sequencing data that needs to be shared, we believe this website will be sufficient for resource sharing purposes. This website will eventually expand to host the data on prioritized variants, and include functionality to enable researchers to query the data using criteria such as genomic region or variant frequency. We note that we have already developed and launched a website for our somatic variant pipeline (funseq.gersteinlab.org, which serves as a preliminary result for eleVAR). Any information linking the variants prioritized by our eleVAR pipeline to the genomes from which they were derived will be removed. Of course, information linking the variants to the individuals will not be included in our released variant dataset.

It is likely that investigators may like to tailor the eleVAR pipeline to their own specific needs or workloads. Thus, we plan to make the source code for eleVAR (as well as the associated modules, such allele DB and LARVA) freely available as tar files hosted at github and at the eleVAR website.

Finally, as we point out the proposed work, eleVAR relies on a complex data context that is derived from many genomic resources. We will make available via the eleVAR website many of these associated annotated files. These include a list of ultra-sensitive regions for TFs and ncRNAs, a list of predicted enhancers, the linkage between enhancers and their target genes, and a set of allelic SNPs and elements.

Narrative

We plan to prioritize rare, germline variants associated with disease for functional impact, using prostate cancer as a test case. We will focus on variants in non-coding regions – a category of variant underrepresented in previous studies. Utilizing a range of genomics data, our goal is to prioritize variants for validation with our eleVAR pipeline.

Abstract

We will investigate potential disease-associated genetic variants in the non-coding regions of the human genome. Recent work in the ENCODE project and in population-scale RNA sequencing has contributed significantly to our knowledge of non-coding elements. Thus, given the focus on coding variation in many previous disease studies, there is much untapped potential in exploring the non-coding variation associated with disease. We plan to prioritize rare, germline non-coding variants for connection to disease, using a generalized framework that we will tune specifically to Prostate Cancer as a test case. Our approach will build upon our existing tool, FunSeq, which prioritizes rare somatic variants in cancer, to create eleVAR – **ele**vating germline **VAR**iants. FunSeq was developed to prioritize somatic variants in regions of the genome depleted of common variants in the general population, based on data from the 1000 Genomes project. eleVAR will use this general principle to analyze germline variations, and build upon it by adding several key features, including: (i) prioritizing variants leading to gain of new transcription-factor (TF) binding sites (in addition to disruption of existing sites), (ii) annotating variants in enhancers and connecting them to target genes, (iii) prioritizing variants highly connected in a variety of biological networks, (iv) annotating variants in non-coding RNAs similarly to those in TF binding sites, and (v) prioritizing variants associated with variable, allele-specific activity. Our second objective is to use eleVAR to prioritize variants in whole genome sequences from the TCGA/ICGC consortium. Our efficient implementation of eleVAR will include a module for updating parameters in response to high throughput experimental data. We will progressively tune and evaluate eleVAR, first using publicly available data, and then using multiple rounds of high throughput experimental characterization of variants occurring specifically in prostate cancer. Our last objective is to functionally validate a subset of variants in details. First, we will identify variants in the 6 representative eleVAR positives and look at their frequency of occurrence in a large prostate cancer cohort using targeted re-sequencing. We will use the CRISPR/Cas system to generate endogenous mutations, determining their effects on target gene expression, cell morphology and tumorigenicity, and TF binding by EMSA and chromatin immunoprecipitation.