Section of gtechr01

D. Approach

D-1 Approach Aim 1 - Convert & extend the FunSeq somatic variant pipeline for germline prioritization

D-1-a Preliminary results for Aim 1

D-1-a-i 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 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\cite{19122651}, prediction of a TF's target genes\cite{22039215}, and new machine learning techniques\cite{19015141}. Furthermore, we developed methods that integrate ChIP-seq, chromatin, conservation, sequence and gene annotation data to identify gene-distal enhancers\cite{20126643}, which we have partially validated\cite{22950945}. 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. \cite{22060676,21177976,21324173,21926158,22955978} We have also constructed regulatory networks for human and model organisms\cite{22955619}\cite{21430782}, and completed many analyses on them (Fig 1)\cite{22125477,21177976,20439753,15145574,14724320,17447836,15372033,19164758,16455753,22955619,22950945,18077332,24092746,23505346,21811232,2160691,21253555}. Furthermore, we conducted large-scale multi-organism regulatory and coexpression network comparisons, along with transcriptome and pseudogene lineage analyses\cite{25164757,25164755,25157146,21253555,25249401}. We also have extensive experience conducting integrated analyses of RNA-Seq datasets generated by the ENCODE, modENCODE, BrainSpan and exRNA consortia\cite{22955616,22955620,21177976,0000001,0000002}. In particular, we developed RSEQtools and IQseq for gene model creation and transcript quantification\cite{21134889,22238592}. We also developed tools that specifically analyze features of ncRNAs, including incRNA and ncVAR for finding and characterizing these elements\cite{21177971, 21596777}.

D-1-a-ii 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 of transcription factors or DNA-binding proteins, and allele-specific expression (ASE)\cite{20567245,20846943}. We have previously developed a tool, AlleleSeq,\cite{21811232} 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\cite{22955619,24092746}\cite{22955620,22955619,24092746}. 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\cite{22955619}. Furthermore, we have constructed a personal diploid genome and transcriptome of NA12878 on\cite{0000003}.

D-1-a-iii 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\cite{21596777,22950945,22955619}. We used metrics, such as diversity and fraction of rare variants, to characterize selection on various classes and subclasses of functional annotations\cite{21596777}. In addition, we have also defined variants that are disruptive to a TF-binding motif in a regulatory region\cite{22955616}. Further studies showed relationships between selection and protein network topology (for instance, quantifying selection in hubs relative to proteins on the network periphery\cite{18077332,23505346}).

In recent studies\cite{24092746,25273974}, 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. It also detects their disruptiveness in TF binding sites (both loss-of and gain-of function events). 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, and it scores somatic recurrent mutations higher than those that are non-recurrent. Using FunSeq, we identified ~100 non-coding candidate drivers in ~90 WGS medulloblastoma, breast and prostate cancer samples \cite{24092746}. ~~We have also applied our method to investigate non-coding mutation patterns in subtypes of gastric cancer\cite{submitted}~~. Drawing on this experience, we are currently co-leading the ICGC PCAWG-2 (analysis of mutations in regulatory regions) group.

HY: CHGNAME ?

HY: BUILD A MODEL ?

D-1-b Research plan for Aim 1

We plan to convert and extend the current FunSeq prototype from its focus on somatic variants to allow the identification of rare germline variants associated with high functional impact (Fig 3). Our new pipeline is called eleVar. It will have several features tailoring it to germline analysis, including 1) identifying functional sites among the conserved regions of the human genome and ncRNA regulatory elements; 2) investigating the allelic elements; and 3) taking into account network connectivity.

D-1-b-i Consistently prioritizing non-coding elements from polymorphism data

In order to define rare variants with highly impactful events, 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\cite{15965027}).

MATH ? FILT ? ANS 2.

We will first update the TF binding non-coding elements from the original FunSeq approach. Here, we will use the better enhancer definition provided by the Epigenome Roadmap \cite{25693563,25533951,25693566}, and more recently from ENCODE. In particular, we will develop a new machine learning framework that utilizes pattern recognition within the signal of various epigenomic features and transcription of enhancer RNA (eRNA) to predict active enhancers across different tissues.

Second, RNA regulatory elements will be added as prioritization features in a way that is consistent with the approach taken for TF binding sites. Specifically, we will mine RNA interactions with proteins/miRNAs from publicly available data, such as CLIP-Seq, CLASH and computational predictions (TargetScan) to create a compendium of biochemical interactions with RNA\cite{25416797, 24297251, 20371350, 23622248, 21909094}. Our initial analyses indicate that some binding sites are even more sensitive to variation than coding sequences. In addition, we will incorporate aspects of RNA 3D-structure. Our initial survey indicates that more rigid RNA structures, such as stems, are under higher selective pressure than other RNA regions, and that those variants that cause a larger free energy change in terms of structure are rarer in human populations. We will define sensitive regions based on folding free energy and folding z-score cutoffs that are enriched for rare genetic variants.

D-1-b-ii Identifying high-impact mutations: breaking & creating motifs

For impactful events at TF binding sites, 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\cite{23512712,24092746,21596777,23348503,23348506,23530248,23887589}. 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.

In a way that is consistent with our means of searching for motif-breaking variants in TF binding sites, we will identify motif-breakers in specific RNA binding motifs. Studies of RNA processing and function have identified key motifs associated with events ranging from RNA splicing to chemical RNA base modifications\cite{18369186}. We have found that intron-exon junctions, polyadenylation sites, and intron lariat structures are much more sensitive to mutation than other genomic regions, particularly for motif-breaking variants. For miRNA/protein bindings sites, we will likewise use the specific binding sites of the microRNAs and whether the respective mutation moves closer to or further from the canonical pattern.

D-1-b-iii Variant prioritization based on allelic activity

Allele-specific variants potentially provide a 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\cite{24037378}, 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 4).

We derive allelic elements by first identifying allelic variants from hundreds of individuals. These individuals will be amassed from The 1000 Genomes Project\cite{23128226}. We will match them with their corresponding RNA-Seq and ChIP-seq experiments from multiple disparate studies, such as gEUVADIS\cite{24037378} and ENCODE\cite{22955616}. Because these separate studies typically have various inconsistencies in terms of tools and parameters used in processing their data, we have to reprocess and harmonize the heterogeneous data and detect allelic variants in a uniform fashion. Also, while the conventional way to detect allelic variants is using the binomial test, previous studies have found that the distributions of the allelic ratios in ChIP-seq and RNA-seq experiments have been empirically observed to give a broader, or an ‘overdispersed’, distribution than a binomial distribution\cite{25223782,20671027,22499706}. To identify and remove problematic "outlier" datasets and to account for overdispersion of read distributions, we will extend our detection pipeline (AlleleSeq) to include the calculation of an overdispersion parameter for each ChIP-seq and RNA-seq dataset; the beta-binomial test (which parametrizes the overdispersion) will be used to detect allelic variants instead of the binomial test.

Subsequently, allelic variants (rare and common) identified across hundreds of genomes can be aggregated into 'allelic genomic elements'. Each element will be assigned an 'allelicity' score based on not only its enrichment of allelic variants within the element (in comparison to accessible variants within the elements and having sufficient coverage to make an allelic activity call), but also across the number of individuals having allelic variants in a consistent allelic direction. The scoring system by element is useful in two ways: (1) it allows continuous ranking of genomic elements based on its allelic impact across multiple individuals (as opposed to defining a threshold to make a binary decision of whether an element is 'allelic') and (2) it enables incorporation of ASE and ASB into the main prioritization scheme; input variants (even those which are rare, but lie in highly-ranked allelic genomic elements) will be upweighted according to their scores.

D-1-b-iv 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\cite{20126643}. 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.

SKL: INSER

We will use the regulatory element-target gene pairs to connect the non-coding variants into a variety of networks -- e.g. regulatory network, metabolic pathways, etc. We will examine their network centralities, such as hubs, bottlenecks and hierarchies, as we know that disruption of highly connected genes or their regulatory elements is more likely to be deleterious\cite{23505346,18077332}. For RNA regulatory elements, we will also use protein/miRNA biochemical interactions to interpret the network context of our variants, using RNA molecules as nodes and RNA-protein and miRNA-RNA interactions as edges. We will prioritize variants that are bound by multiple factors, and those within whole RNAs that are bound by many RNA-binding proteins.

D-1-b-v We will use a unified weighted scoring scheme for combining all eleVAR features to prioritize variants

To integrate the various features mentioned above, we plan to elaborate the weighting system in FunSeq.\cite{25273974}. 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, such as sentitive region overlap, ultraconserve region overlap, and HOT region overlap, we calculate the probability p_d that it overlaps with common polymorphisms. We then calculate the information content to denote the value of discret e features $s_d = 1 + p_d * log_2 p_d + (1 - p_d) * log_2(1 - p_d) + \theta_d$, where $\theta_d \sim N(0, \sigma)$ and can be used for score optimization.

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 , such as motif gain, motif break and GERP etc, 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 $s_c^{v_c} = 1 + p_c^{v_c} * log_2 p_c^{v_c} + (1 - p_c^{v_c}) * log_2(1 - p_c^{v_c})$. We then fit a smoothing curve and estimate parameters θ_c's according to empirical distribution $S_c \sim v_c$.
The eleVAR score (eS) is calculated by summing up the values of all its features $\boldsymbol{eS} = \sum_d s_d + \sum_c s_c^{v_c} = \sum s(\theta)$. We will also consider the feature dependency structure when calculating the scores (e.g., removing redundant features or performing dimension reduction techniques).

~~SKL~~
Chg
?

D-2-b-i Statistical framework for parameter tuning using Bayesian updates

The initial feature parameter $\boldsymbol{\theta}$ $(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \ldots, \boldsymbol{\theta}_m)$ (given m number of features) assigned in D-1-b-v will be further optimized with newly available "gold standard" datasets. We plan to tune these parameters using an incremental Bayesian learning strategy. For a variant A, given eleVAR score eS (equation 3 in D-1-b-v), the probability that v is functional ($y_v = 1$ designates a positive result, whereas $y_v = 0$ denotes a negative result) follows a logistic function $P(y_i = 1|\boldsymbol{\theta}, eS) = \frac{1}{1 + \exp(-k * (eS - a))} = \frac{1}{1 + \exp\ (-k * (\sum S(\boldsymbol{\theta}) - \alpha)}$ $(k, a$ are scaling parameters). To update θ with training data $\boldsymbol{Y}$, we implement Bayes' rule: $P(\boldsymbol{\theta}|\boldsymbol{Y}, \boldsymbol{eS}) \propto P(\boldsymbol{Y}|\boldsymbol{\theta}, \mathbf{eS})P(\boldsymbol{\theta})$.
The likelihood ratio is defined as: $Q = \frac{L(Y|\theta_{proposed})}{L(Y|\theta_{current})}$, and then MCMC (Monte Carlo Markov Chain) will be used to find the most probable $\boldsymbol{\theta}$. The updated $\boldsymbol{\theta}$ will then be used as tuned parameters in eleVAR to prioritize variants. The procedure will be iterated in several rounds. In the first round of tuning, feature weights obtained in D-1-b-v will be used to construct priors $P(\boldsymbol{\theta})$. In subsequent rounds, the updated weights will be set as new priors.

D-2-b-iii Generating an initial list of prioritized variants & then running them through eleVAR

We'll get rare variants from 1000G Phase 3. [[MG: remove to PCAWG]][[SKL: to cut this part? ~~D-1 has already used 1KG common SNP to train parameters, the original text focus on cancer and tissue specific stuff, which is nothing with current context, So cut this part and go to D-2-b-iv?]]~~

~~We will run eleVAR on the rare variants resulting from our variant calling on 2658 PCAWG and on the predictive compendium whole-genome sequences.~~

D-2-b-iv Round 1 of 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. These datasets include known disease-causing mutations from molecular studies, high throughput reporter assays on enhancer activities ~~and recurrence of cancer rare mutations in the region of interest involving germline and potentially somatic variants.~~

The Human Gene Mutation Database (HGMD)\cite{12754702} and ClinVar\cite{24234437} catalogue large numbers of regulatory disease-causing mutations discovered in molecular studies. 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\cite{23129659} to assay over 1,000 single- and double-nucleotide mutations in promoter regions. Kheradpour et al.\cite{23512712} used MPRA to test variants affecting regulatory motifs in over 2,000 human enhancers. We will utilize these datasets to perform comparisons with other variant prioritization methods, such as CADD\cite{24487276}, to obtain a preliminary evaluation of method performance. We will then tune our parameters using the scheme described above.

TO 4 #~~4~~ 7

D-2-b-v Round 2 of tuning using high-throughput experiments done in this project

We expect an average of ~40K rare germline variants per genome\cite{23128226}. Since they rarely recur at the exact same position, we anticipate a prioritized list of ~8M variants (=40K * 250 genomes, based on the the expected size of the prostate compendium). We will select 500 functional regions of appreciable size that contain highly ranked variants. Assuming ~8M variants are distributed evenly across the human genome, taking an average element size of 3kb, the number of variants per element will be ~4. Variants on the same element are expected to have different functional impacts. For each element, we will prioritize at least one of these variants to be of high impact, and the remaining variants to be of a lower impact. Specifically, we will have a total of 1000 variants (500 with a high impact and 500 with a low impact). Subsequent tuning and refinement of the eleVAR parameters will be based on further experimental characterization of these 1000 variants (500 highly prioritized and 500 lowly, respectively). We will validate these variants through functional genomic screens using the [[change cloneseq]] [[SKL:Done]]**STRO-seq** technology coupled with luciferase reporter assays. Overall, this refinement will be accomplished in two rounds, each round per year, as detailed in Aim 3 and the timeline (Fig 6). Finally, during the last year of the proposed work, we will perform a careful assessment of our model. We will again prioritize our full list of variants and select a final set of 200 top ranked variants for an unbiased validation. This will allow us to construct a precise ROC curve in order to measure the accuracy of our predictions.

CONFIRMED #4
UPDATE