Evaluating the functional impact of coding and non-coding somatic mutations over multiple scales

Systems biology U01 grant
June 11, 2017

Background and significance: 1,093
Aim 1 (prior experience): 1,023
Aim 1 (approach): 2,576
Aim 2: 3,332
Aim 3: 1,447

3.5K
5 PG

A. SIGNIFICANCE

A.1. Different Scales. The typical cancer has thousands of somatic variants. These manifest their effects on different scales. At the smallest scale is direct effect molecular activity such as the binding of a transcription factor or the transcription of a downstream gene (often dubbed the "molecular endophenotype" \cite{24748924, need better ref}). Cancer manifests itself on a cellular level in terms of the phenotype of the cells -- e.g., growth or invasiveness, the latter related to metastasis. Finally, it also manifests itself on an overall organismic level in terms of, obviously, the cancer phenotype, but more subtly in terms of the severity of the cancer.

The extent to which variant effects take place at the levels of molecular activity propagates to the cellular phenotype, and organismal presentation is unclear.

A.2. Systems & Networks. The coupling between the individual variant, its effect on molecular activity, cellular phenotype, and organismal development, is a systems effect. With many variants, genes are often connected in both regulatory and interaction networks. We endeavor to probe these connections here. In particular, many proteins carry out diverse functions through interacting with other proteins [3]. Recent studies have been conducted on genetic coding mutations in the context of the human interactome network [4-7], where on average, a functionally active protein interacts with >5 other protein partners. We will leverage our experience to deploy a novel approach that systematically uses several agnostic functional assays in parallel. This approach serves as a paradigm to prioritize coding variants and provides important insights into mutation mechanisms of interaction from a systems biology perspective.

A.3. Evaluation of Coding and Noncoding Variants. Conceptually, both coding and noncoding variants may vary in their degree of impact on cancer development or protein formation and function. Numerically, the overwhelming bulk of variants in cancer genomes are non-coding (usually by a factor of 50 to 100) \cite{26781813}. Historically, there has been an emphasis towards studying coding variants due to the functional significance of protein coding regions. However, as noncoding alterations constitute the majority of disease-associated variants [1], further study of non-coding regions may also be critical to a better understanding of cancer biology. Accordingly, we will consider a combination of coding and noncoding variants. Moreover, a wealth of non-coding information is available due to advances in sequencing technologies and efforts by consortia like ENCODE and 1000 Genomes \cite{22955616, 20981092}.

A.4. Weaker effects. Non-coding variants traditionally have been thought to have weaker effects than coding ones -- not disabling a gene or creating a new binding site in one but more subtly affecting regulation. These may come into play in the development of weaker drivers which may have smaller effects on cancer. There has been recent work on these of late. In particular, recent studies \cite{26456849} \cite{23388632} suggest that certain passenger mutations described as "mini-drivers" may have a weak effect on tumor cell fitness and in turn promote or inhibit tumor growth. From a tumor fitness perspective, three categories can thus emerge: positively-selected driver variants, neutrally-selected passenger variants, and negatively-selected mini-driver variants.

A.5. Application to prostate cancer. Prostate cancer is a particularly tractable system for us to focus on for a number of reasons. First of all, as we described, we have much preliminary background working on this specific cancer and deep connections with the cancer SPORE grant.[[add ref]] Also, prostate cancer is highly heterogeneous, displaying very different phenotypes, from a highly indolent, almost notice less disease, to a very aggressive condition. These different presentations may be coupled to systems-wide effects.

Significant efforts have been made to study genetic and environmental causes of this cancer type, but major leaps forward are still needed to develop a more complete etiology of this disease that affects XXX million men worldwide. Along with other major factors associated with prostate cancer such as the hormonal action of androgens and estrogens [8], more than 70 genetic susceptibility variants have been identified [9]. Suspected loci are continuously being discovered using GWAS studies [10] and genotyping arrays [11]. Such variants increase the predictability of the disease and have been associated with altering the expression levels of several genes. Some of the most well-known genes associated with prostate cancer are TP53 and RB1 \cite{28586335, add more refs?}. These genes are both tumor suppressors, and their alteration is associated with poorly differentiated tumors of prostate and other cancers, which tend to be more aggressive\cite{28586335}. We will conduct focused investigations of coding and noncoding variants associated with these genes and their molecular subnetworks.

A.6. Indolent versus aggressive. One of the most interesting questions about prostate cancer is whether if can detect the overall aggressiveness of the disease from its molecular mutation profile as this has direct implications for treatment.[[ref]]

REW

B. INNOVATION

Our mathematical model, its multi-tiered cutting-edge biological validation in concert and each individually, and the real-time Bayesian update of the former with the latter are fresh, exciting contributions to the field.

B.1 Overall Framework. We believe our overall approach is highly innovative in that we have assembled a diverse team of investigators and are probing prostate cancer on many levels, from clinical outcomes to a more cellular, systems-wide experiments, to large-scale molecular experiments to computational prioritization on a variety of scales.

B.2 Aim 1 - Mathematical Model. The specific mathematical model that we are developing is innovative for a number of reasons. First of all, it encompasses a wide range of genomic features. Second, of all, it combines information from both the molecular, nucleotide-level scale (biochemical/biophysical, evolutionary, and network) with information about recurrence and whole-organism disease phenotype. Second, we provide an innovative scheme to update our model in a Bayesian framework using large-scale experimental data. The update and the validation will lead to a more accurate and usable model.

HY
MORE

B.3 Aim 2 - High-throughput Molecular Experiments. eSTARR-seq: this unique barcoding approach allows direct quantification of enhancer activity, with 40-fold increase in sequencing efficiency compared with traditional STARR-seq

InPOINT: this unique technology directly examines the biochemical consequences of coding variants on protein stability and interactions

HY
TRUE
2

B.3 Aim 3 - Cellular Assays. CRISPR: This genomic editing breakthrough technology can build a cellular variants impact evaluation model to introduce targeted mutation in coding and noncoding regions from normal prostate cell lines, which will grow in prostate organoid to investigate tumor progression effect. Organoid technology: This technique, successfully deployed differs from traditional cell culture by maintaining cancer cells in three-dimensional (3D) cultures. Benign and cancer cells that are grown in 3D retain cell-cell and cell-matrix interactions that more closely resemble those of the original tumor compared to cells grown in two dimensions on plastic.

C. APPROACH

C.1. AIM 1 Computational prioritization of coding and non-coding somatic mutation We will first prioritize both coding and noncoding prostate cancer variants. ~~This prioritization will be used to identify variants to be~~ investigated using subsequent assays of molecular, cellular, and organoid-level phenotypes. These assays will

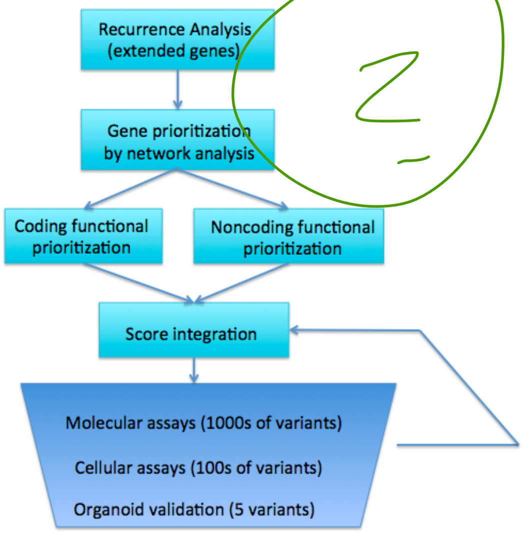
HY 2

simultaneously validate candidate oncogenic variants and refine tools to predict impactful variants (Figure 1). These efforts leverage our extensive experience in both variant prioritization and cancer genome analysis.

C.1.A. Prior experience for variant prioritization

LATER

C.1.A.1. Experience in background mutation rate estimation and recurrence analysis. A major method to search for driver variants is to find genes or regions of the genome that are highly enriched for mutations. However, this search can be confounded by the fact that different regions of the genome have different mutation rates. Moreover, great mutation heterogeneity and potential correlations between neighboring sites give rise to substantial overdispersion in mutation counts, which complicates background rate estimation. We developed a computational framework called LARVA, which integrates variants with a set of noncoding functional elements, modeling the mutation counts of the elements with a beta-binomial distribution to handle overdispersion \cite{26304545}. Importantly, this method incorporates regional genomic features such as replication timing to better estimate local mutation rates and find mutational hotspots. Applying LARVA to 760 whole-genome tumor sequences shows that it identifies well-known noncoding drivers, such as mutations in the TERT promoter, in addition to uncovering new potential noncoding driver regions.


Recurrence Analysis (extended genes)
Gene prioritization by network analysis
Coding functional prioritization
Noncoding functional prioritization
Score integration
Molecular assays (1000s of variants)
Cellular assays (100s of variants)
Organoid validation (5 variants)

2

Figure 1. Overall variant prioritization workflow

C.1.A.2. Experience prioritizing protein-coding variants. We have developed a number of tools that identify deleterious protein-coding variants. Our Variant Annotation Tool (VAT) is a utility that characterizes variants according to the genes and transcript isoforms affected, and any amino acids changes that may result \cite {22743228}. Building upon this work, Analysis of Loss of Function Transcripts (ALoFT) is a software pipeline we developed to predict which mutations will cause gene loss of function and whether loss of one or both copies of a gene is necessary to observe this effect. Assessing the functional impact of loss-of-function variants is one purpose of our netSNP tool that integrates networks of protein-protein interaction, transcription factor binding, and metabolic pathways to build a classifier that distinguishes essential genes (Fig 1) \cite{23505346}. The application of this tool to cancer genomes shows an enrichment for predicted loss of function mutations in known cancer-associated genes. STRESS is a tool we built to identify mutations that might affect allosteric hotspots in proteins, which can be key to protein function \cite{27066750}. Similarly, our Frustration tool uses calculations of localized structural frustration to identify key functional protein regions that may be altered by genetic variants \cite{27915290}. Finally, our Intensification tool searches for deleterious mutations particularly within repeat regions of proteins \cite{27939289}.

SHORTEN ~30%

UP

1000G

C.1.A.3. Experience in noncoding genome analysis. Our expertise in prioritizing noncoding DNA variants is built on our experience analyzing a wide variety of genomic assays. Much of this work has been in connection with the ENCODE and modENCODE consortia \cite{22955616, 25164757, 22955619, 21177976}. We have developed widely used tools to identify ChIP-Seq peaks \cite{19122651, MUSIC}, perform RNA-Seq quantification \cite{21134889, 22238592}, and identify new noncoding transcripts and categorize them according to function \cite{21177971, 25164757}. Our tool to predict enhancer regions \cite{22950945} has undergone functional validation of its predictions \cite{#58 from ncvarg grant, find PMID}. We have further linked enhancers to target genes \cite{25273974}, and have developed related tools to process HiC data \cite{28369339, http://biorxiv.org/content/early/2016/12/29/097345}. This work highlights chromosome conformations that can aid enhancer-target linkage inference. In addition to identifying, quantifying, and linking noncoding genomic elements, we have multiple linear and nonlinear models that use epigenetic signals to predict gene expression \cite{22955978, 21926158, 21324173}. Moreover, we have extensive experience incorporating genomic data into networks to help explain gene regulation and to identify key regulators \cite{22955619, 25249401, }.

C.1.A.4. Experience in allelic analysis. We have also made focused investigations of allele-specific activity in the genome, which can provide a direct readout of the effects of an allele-specific variant (ASV). We developed the AlleleSeq pipeline to quantify allele-specific expression \cite{21811232}. More recently, we conducted a study of allele-specific activity from RNA-Seq and ChIP-Seq experiments conducted on 1000 Genomes Project \cite{23128226, 27089393} individuals, including data from the gEUVADIS \cite{24037378} and ENCODE \cite{22955616} projects. After uniformly reprocessing all data, we detected ASVs using a beta-binomial test to correct for overdispersion. Since most ASVs are rare variants, we also combined the effects of many variants to assign allelicity scores to genomic elements, indicating that these elements are particularly sensitive to mutations.

HEADING

C.1.A.5. Experience in noncoding variant prioritization. We have completed extensive analysis of patterns of variation in noncoding regions, along with their coding targets[90,95,114]. 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 noncoding mutation to target genes, and prioritizes such variants based on scaled network connectivity. It identifies deleterious variants in many noncoding functional elements, including TF binding sites, enhancer elements, and regions of open chromatin corresponding to DNase I hypersensitive sites. By integrating data from large-scale resources (including ENCODE and the 1000 Genomes Project) with cancer genomics data, our method is able to prioritize known TERT promoter driver mutations.

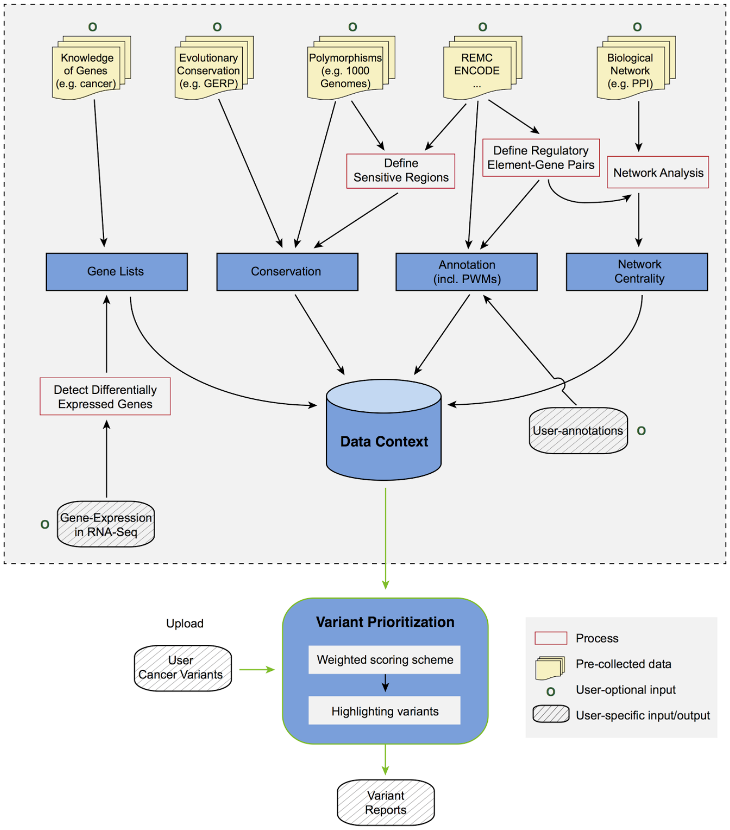


Figure 2. FunSeq2 workflow

AS COLEAD

C.1.A.6. Experience in genomics and cancer genome analysis consortia. We have extensive experience in analysis of cancer genomes through our participation in The Cancer Genome Atlas (TCGA) and Pancancer Analysis of Whole Genomes (PCAWG) consortium projects. We participated in the TCGA consortium studies of prostate \cite{26544944} and kidney \cite{26536169} cancers and recently conducted a detailed investigation of the noncoding variants in TCGA kidney papillary cancer samples \cite{28358873}. We developed tools for somatic variant calling \cite{26381235}. We used TCGA RNA-Seq data extensively in the development and application of seveal other tools \cite{Loregic, DREISS, 25884877}. We are currently leading the PCAWG group investigating the impact of so-called passenger mutations on cancer development, progression, and prognosis. We are also ~~conducting a study integrating ENCODE data to provide a comprehensive resource for cancer to interpret patient cohort data such as expression and somatic mutation profiles~~.

C.1.B. Research plan for variant prioritization

C.1.B.1. Identification of recurrently mutated elements & genes

FLIP

C.1.B.1.a. Tools for background mutation rate estimation and recurrence analysis
To identify genes whose mutation is important to the development of prostate cancer, we will search for genes that are recurrently mutated in prostate cancer patients. We will do this using the compact annotation described above, which incorporates both coding and noncoding elements associated with a given gene (see below).

We also propose a Negative binomial regression based Integrative Method for mutation Burden analysis (NIMBus). This analysis will treat mutation rates from different individuals as random variables with an underlying a gamma distribution. Pooled mutation counts from a heterogeneous population serve as a negative binomial distribution to handle overdispersion. Furthermore, to capture the effect of covariates, NIMBus integrates extensive features in all available tissues from Roadmap Epigenomics Mapping Consortium (REMC) and the Encyclopedia of DNA Elements (ENCODE) project. The result of this data integration is a covariate matrix that predicts local mutation rate with high precision through regression. In addition, customization of the most comprehensive noncoding annotations from ENCODE facilitate the interpretation of results. This integrative approach will enable us to effectively pinpoint mutation hotspots associated with disease progression. This genome-wide search for disproportionately burdened regions will be complemented by use of our LARVA software for identification of recurrent mutation affecting non-coding elements.

MENTION COMPACT EARLIER HOW REL TO LARVA?

~~Our search for significant genomic regions across the breadth of the genome, will then be complemented by detailed examination of the molecular functional impact of individual variants. This examination of the molecular functional impact of individual variants will include detailed follow-up examination of any genomic regions of significance identified through genome-wide recurrence and burdening analysis. We will examine individual variants in both coding and noncoding regions.~~

MOVE UP

C.1.B.1.b. Definition of a compact annotation for variant analysis

To perform combined recurrence analysis of coding and noncoding elements, we will take non-coding regulatory regions and link them to genes by constructing "extended gene neighborhoods". We will first identify a compact list of enhancers through a purpose built ensemble method. This ensemble method integrates ChIP-seq, DNase-seq, and STARR-seq into a pipeline for enhancer candidate identification based on pattern recognition. Enhancer-target linkages will be predicted using the Joint Effect of Multiple Enhancers (JEME) method of Cao *et al*. (under review). Subsequent filtering of these predictions will be performed using high-resolution Hi-C experiments. We will also extract cis-acting TF and RBP binding sites and incorporate them into these extended genes. Similar to exonic regions within genes, we will annotate a set of discrete regions that potentially affect gene expression. This unified annotation will enable joint evaluation of the mutational signals from distributed yet biologically relevant genomic regions.

C.1.B.2. Variant prioritization by molecular disruption

HEADING 2

MOTIFS & POWER

C.1.B.2.a. Functional prioritization of coding mutations

Once we have identified putative driver genes through a combination of recurrence and biological network analysis, we will score the functional importance of mutations that overlap the coding regions of these genes. We will use our VAT and ALoFT tools to identify mutations that may completely inactivate copies of genes. For potentially impactful variants that do not fully eliminate gene function, we will combine GERP score - a measure of evolutionary conservation - and FunSeq2 score, an ensemble method that combines scores from many tools to score the functional impact of coding variants \cite{ 24453961} In addition, for proteins with known structures, we will apply our STRESS \cite{27066750} and Frustration \cite{27915290} tools to search for allosteric hotspots and sites of localized structural frustration, respectively. We will also use our Intensification tool to provide mutation impact scores within protein repeat regions \cite{27939289}.

SHORTEN

SHORTEN

C.1.B.2.b. Functional prioritization of noncoding mutations

FunSeq and FunSeq2 allow us to score mutations based on predicted molecular functional impact. Variants with high FunSeq scores are predicted to be functionally impactful variants. These high scoring variants tend to be located in functionally significant noncoding domains, and may correspond to undiscovered drivers (both strong & weak) as well as passenger variants that decrease tumor cell fitness. Conversely, common variations tend to arise in functionally unimportant regions due to constraint by selective pressure. Thus, genomic features that are enriched with common polymorphisms are less likely to contribute to the deleteriousness of variants and so receive lower scores.

We will expand the scoring system of FunSeq\cite{24092746} and Funseq2\cite{25273974} in order to integrate the additional variant attributes we measure. In general, features can be classified as discrete (e.g., either within or outside of a given functional annotation) or continuous (e.g., the PWM change in 'motif-breaking'). We will weigh these two sets of features using different strategies. For each discrete feature, we will calculate the probability that it overlaps with common polymorphisms. We will 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 first 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 $s_c^{v_c} = 1 + p_c^{v_c} * log_2 p_c^{v_c} + (1 - p_c^{v_c}) * log_2(1 - p_c^{v_c})$.

HOW INTEG LARVA

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

MOVE DOWN

C.1.B.2.c. Identification of key regulators using TF network analysis

We will ~~then~~ investigate the global topology of the transcriptional regulation network by comparing the inbound and outbound edges of each transcription factor (TF). The level of a TF within the network hierarchy reflects the extent to which it directly regulates expression of other TFs \cite{25880651}. TF rewiring (i.e., target changing) may help to identify cancer-associated deregulation when comparing the common regulators in approximately matched tumor and normal regulatory networks. Our rewiring analysis not only considers direct connections associated with a given TF, but also the whole neighborhood of connections with which a TF associates. Through use of both TF membership and topic models, we are able to build a mixed-membership model to look more abstractly at local gene neighborhoods to re-rank the TFs.

CUT 50%
NO REWIRE

C.1.B.3. Emphasis on variation affecting *TP53* and *RB1*

Given the high frequency of somatic variation affecting of TP53 and RB1 tumor suppressor genes across cancer types, we will focus on the characterization and prioritization of variants affecting these genes. Recurrent mutations of TP53 and RB1 are particularly common in poorly differentiated neuroendocrine tumors of prostatic origin \cite{28586335}. Transdifferentiation of prostatic adenocarcinoma into a neuroendocrine phenotype may occur concomitant with resistance to chemotherapeutic agents \cite{28411207}. This finding underscores the importance of detailed examination of variation affecting RB1 and TP53. An understanding of mechanisms of gene loss may lead to improved therapeutic options for patients that develop resistance to chemotherapy. Our expertise in understanding the effect of variation affecting noncoding regions, and the our ability to predict the consequence of variation affecting networks of genes, and extended networks of genetic regulatory elements, will be used to provide an in-depth understanding how TP53 and RB1 levels may be affected by variants that do not affect their coding sequence.

MOVE TO INTRO

C.1.B.4. Updating model parameters following experimental validation

Let $\boldsymbol{\theta}^{(0)} = (\theta_1^{(0)}, \theta_2^{(0)}, \dots, \theta_m^{(0)})$ represent an initial feature parameters chosen at random, 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 2 and 3. Because of the high throughput of our molecular activity assays (eSTARR-seq and InPOINT), our strategy is to implement for the first time an iterative learning scheme consisting of three stages: 1) initial learning, 2) real-time experimental parameter optimization, and 3) final assessment.

(SEE LATER)

In the first stage, we will randomly select ~500 candidate driver genes as defined by recurrence analysis, PCAWG and TCGA. We will first generate the wild-type and mutant clones of these genes and promoters using Clone-seq. Then we will select 2 coding variants from the coding region and 2 non-coding variants from the promoter region of each gene and generate all ~2,000 variant clones through Clone-seq. The effect of these variants on coding and non-coding variants will be quantified by the InPOINT and eSTARR-seq pipelines respectively. Starting from the initial tuning of $\boldsymbol{\theta}^{(0)}$, we will update these tunings according to the results of

MORE GEN

~2000 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. The expectation of y_v can be predicted through a logistic regression: $\mathrm{logit}(P(y_v = 1)) = -k * (\boldsymbol{RS_v} - a) = -k * \left(\sum_m \theta_m * s_{v,m} - a\right)$ $(k, a$ are scaling parameters). To update $\boldsymbol{\theta}^{(\mathbf{0})}$ 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 Chain Markov Carlo) sampling to search over the parameter space and find the most probable $\boldsymbol{\theta}^{(\mathbf{1})}$. We will predict the functional impact of all noncoding variants genome-wide, $P\left(y_v = 1|\boldsymbol{\theta}^{(\mathbf{1})}\right)$.

~~In the third stage of final assessment, we will select xxx variants (xxx with predicted high impact, xxx with medium impact, and xxx with low impact) on previously cloned candidate driver genes. We will measure their impact on cell growth and migration activities quantitatively through xx-seq.~~

~~We will build similar models for both molecular and cellular activity assays, with the exception that some gene-centric features, e.g. protein-protein interaction network degree, will be held out of noncoding activity prioritization.~~

C.2 AIM 2 High-throughput *in vitro* quantification of molecular phenotypes of ~2500 non-coding and ~1500 coding mutations

In Aim 2, we will take variants prioritized using our models in Aim 1 and investigate their molecular activity in eSTARR-Seq assays for noncoding variants and using our InPOINT pipeline for coding variants. Results of these assays will enable model tuning and aid selection of candidates for assays of cellular phenotypes in Aim 3.

C.2.A. Preliminary results

C.2.A.1. The importance of investigating functional relevance of coding variants through protein interactome networks

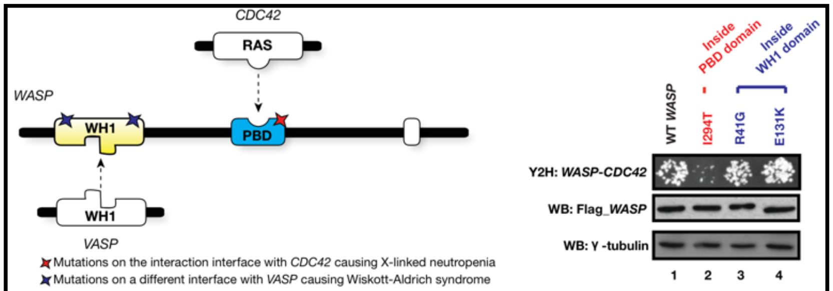

CDC42
RAS
WASP
WH1
PBD
WH1
VASP
Inside PBD domain
Inside WH1 domain
WT WASP
I294T
R41G
E131K
Y2H: WASP-CDC42
WB: Flag_WASP
WB: Y -tubulin
1
2
3
4
Mutations on the interaction interface with CDC42 causing X-linked neutropenia
Mutations on a different interface with VASP causing Wiskott-Aldrich syndrome

Fig 1. Illustration of *WASP*'s interaction interfaces with *CDC42* and *VASP* and effects on the *WASP-CDC42* interaction by mutations on *WASP*. Mutations causing two distinct diseases are located on two separate interaction interfaces and disrupt different interactions as described in our ***Nature Biotechnology*** paper.

An increasingly accepted view of the cell is that of a complex network of interacting macromolecules and metabolites, sometimes referred to as the "interactome network"[1]. In particular, protein-protein interactome networks are of great importance because most proteins carry out their functions by interacting with other proteins[1,2]. More importantly, many proteins are pleiotropic and carry out diverse functions through interacting with different proteins[3]. On average, a protein interacts with >5 other protein partners in the human interactome network. Recently, studies have been conducted on genetic coding mutations in the context of the human interactome network[4-7]. However, our approach is novel in that we systematically use several agnostic functional assays in parallel.

Previously, as described in *Nature Biotechnology*, *Science*, and *AJHG*[8-10], improved upon here in preliminary results (see **c.1.1.3**, **c.2.1.2**, and **c.2.1.4**), our team has successfully used our high-throughput InPOINT pipeline to screen >2,000 coding genetic variants and successfully identified many deleterious genetic mutations, for example, in the Wiskott-Aldrich Syndrome Protein (WASP, see **Fig. 1**). This strategy also provided important insights into mutation mechanisms, in particular

MOVE TO SIGNIF

that many coding mutations only affect a subset of specific interactions, rather than all interactions, and that mutations in the same protein disrupting different protein-protein interactions often lead to clinically distinct disorders[10-13]. **Overall, our InPOINT screen both effectively nominates candidate mutations and gives insights into specific mechanisms to be tested in follow up confirmatory assays.**

C.2.A.2 Our site-directed mutagenesis Clone-seq pipeline is unique

Our recently-published Clone-seq pipeline allows massively-parallel site-directed mutagenesis to generate ***one and only one specific*** mutation per DNA molecule for ***thousands*** of genes/TREs (enhancers and promoters). We have used our Clone-seq pipelines to generate thousands of gene/enhancer WT and mutant clones with an average length of ~2kb. We will have no problem cloning enhancers (up to 4kb) and their mutations in their entirety. Clone-seq is entirely different from previously described random mutagenesis approaches[50-53]: each mutant clone has a separate stock with one and only one pre-defined mutation. Finally, we implemented a smart-pooling strategy and a customized variant-calling algorithm such that we can fully sequence each mutant clone in its entirety and ensure that there are no other unwanted mutations introduce on clones used in all downstream experiments (e.g., iSTARR-seq, InPOINT, or other *in vivo* functional assays).

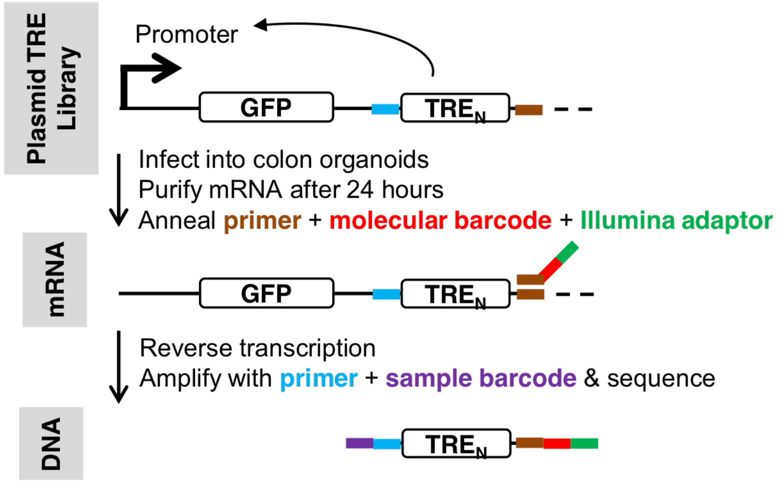

Plasmid TRE Library
Promoter
GFP
TRE_N
Infect into colon organoids
Purify mRNA after 24 hours
Anneal primer + molecular barcode + Illumina adaptor
mRNA
GFP
TRE_N
Reverse transcription
Amplify with primer + sample barcode & sequence
DNA
TRE_N

Fig 2. Our modified iSTARR-seq pipeline. The random molecular barcodes (red) uniquely label each mRNA molecule produced before the amplification step during the sequencing library preparation, enabling us to quantitatively determine the enhancer

C.2.A.3. iSTARR-seq: highly parallel transcriptional readout of candidate regulatory variants

STARR-seq (self-transcribing active regulatory region-sequencing) is a recently-established method that can identify enhancer elements genome-wide[14]. 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 the original STARR-seq protocol ***cannot*** be directly used to measure enhancer activities from a clonal library of WT and mutant enhancer elements, where each element has one and only one clone with defined boundaries, as is the case for our proposed research. Furthermore, >98% of sequencing reads are discarded in STARR-seq because multiple mRNA molecules are often produced from a single unique DNA fragment (see Supplemental Figure 2E of Arnold et al[14]). To circumvent these difficulties, we developed the chromosome-integrated STARR-seq (**iSTARR-seq**) transcriptional readout assay to incorporate a unique molecular barcode to the cDNA of each mRNA molecular produced at the reverse transcription step, allowing direct quantification of enhancer activity for each individual enhancer by counting RNA sequence reads with unique molecular barcodes (**Fig. 2**). In our preliminary study (**c.1.1.4**), >80% of the reads were used for enhancer activity quantification (>40-fold increase in sequencing efficiency). In summary, these improvements will significantly simplify high-throughput studies of candidate enhancer sequences, and increase assay sensitivity compared with the original STARR-seq protocol.

b.4. Our high-throughput InPOINT pipeline that directly examines the biochemical consequences of coding variants on protein stability and interactions is innovative

As described in our previous publications (e.g., *Nature Biotechnology*, *Science*, *PLoS Genetics* and *AJHG*[8-10,12]), our InPOINT pipeline incorporate six high-throughput approaches: Clone-seq (to generate specific mutant clones), GFP (to examine SNP's impact on protein stability), and four orthogonal interaction assays (PCA, LUMIER to examine SNP's impact on specific protein-protein interactions).

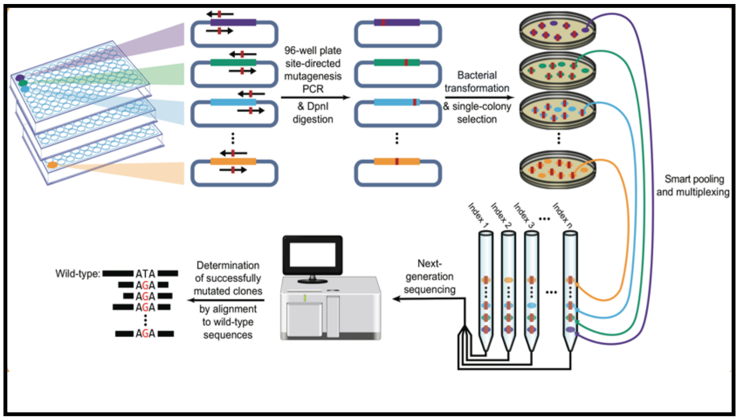
C.2.B. APPROACH

2

c.2. Specific Aim 2. High-throughput *in vitro* quantification of molecular phenotypes of ~2500 non-coding and ~1500 coding mutations.

c.2.1. Preliminary Studies

c.2.1.1. Performance, throughput, and cost of our Clone-seq pipeline. Clone-seq is currently the highest-throughput site-directed mutagenesis pipeline for generating thousands of targeted mutations on many genes. Clone-seq is entirely different from previously described random mutagenesis approaches[50-53]: each mutant clone has a separate stock with one and only one pre-defined mutation. Other methods, such as Dial-out PCR[15], are not comparable because it can only generate clones of short fragments limited by the Illumina read length. In Clone-seq, we routinely clone genes of length >4 kb; each clone is fully sequence-verified at part of the pipeline (**Fig. 5**) to ensure it has one and only one pre-defined mutation. Every step of Clone-seq has been significantly optimized for high-throughput operations. We have also implemented customized variant calling software because existing pipelines (e.g., GATK[16]) cannot be applied due to our pooling strategy[12]. This customized variant calling software allows us to carefully examine whether other unwanted mutations have been inadvertently introduced during PCR-mutagenesis throughout the entire clone.


96-well plate site-directed mutagenesis PCR
& DpnI digestion
Bacterial transformation
& single-colony selection
Smart pooling and multiplexing
Index 1
Index 2
Index 3
...
Index n
Next-generation sequencing
Determination of successfully mutated clones
by alignment to wild-type sequences
Wild-type:
ATA
AGA
AGA
AGA
AGA

The Clone-seq pipeline can easily be adapted to clone WT TREs and genes. To date, we have used the Clone-seq pipeline[12] to successfully generate 678 WT TRE clones and 4,026 mutant clones on 2,438 TREs/genes. The results confirm the scalability, accuracy, and throughput of our Clone-seq pipeline. We are confident that this approach can successfully generate all WT and mutant clones as proposed. ***c.2.1.2. We have successfully implemented our iSTARR-seq assay to quantitatively measure enhancer activities of 678 TREs and their mutations.*** 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-iSTARR-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[14] as controls. Additionally, we used Clone-seq to generate WT and mutant clones for 678 TREs. We cloned all WT and mutant TREs in pDEST-iSTARR-dCP by Gateway LR reaction and quantified their enhancer activity through our iSTARR-seq assay (**Fig. 6a**). 49 of the 346 (14.2%) TRE mutations examined show significantly lower enhancer activities measured by iSTARR-seq as compared to their corresponding WT TREs. Additionally, all five control fragments were also cloned into pGL4.23-DEST-dCP vector and their enhancer activity was also confirmed by the dual luciferase assay. Both experiments (**Fig. 6bc**) successfully replicated the data published in the original STARR-seq paper[14]. Thus, the Gateway-compatible iSTARR-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 iSTARR-seq vectors representing the two major classes of core promoters[17]: one that is responsive to developmental enhancers (pDEST-hSTARR-dCP) and one responsive to housekeeping enhancers (pDEST-hSTARR-hkCP).

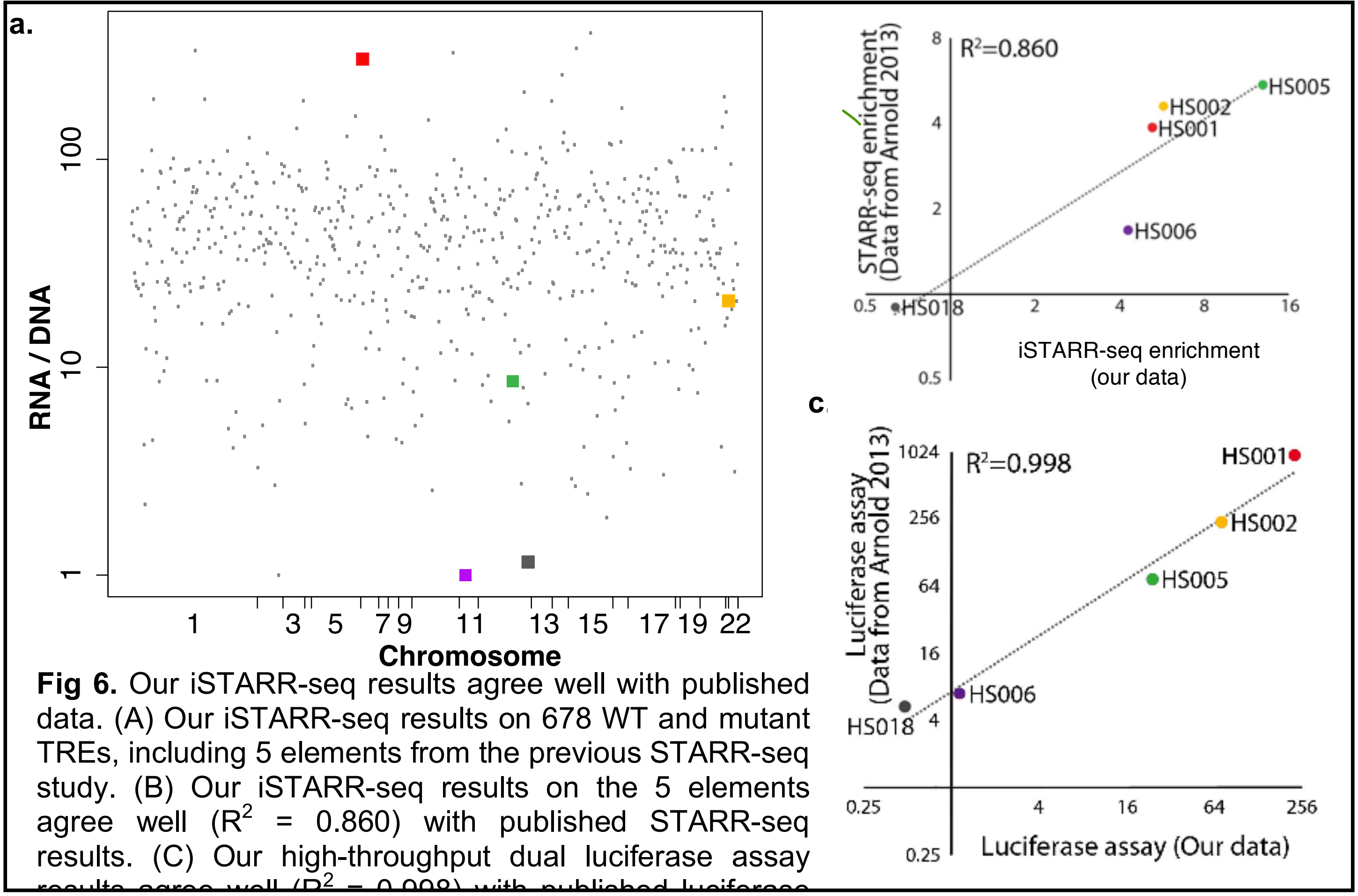


Fig 6. Our iSTARR-seq results agree well with published data. (A) Our iSTARR-seq results on 678 WT and mutant TREs, including 5 elements from the previous STARR-seq study. (B) Our iSTARR-seq results on the 5 elements agree well (R^2 = 0.860) with published STARR-seq results. (C) Our high-throughput dual luciferase assay results agree well (R^2 = 0.998) with published luciferase

c.2.1.2. Using our high-throughput InPOINT pipeline (GFP assay) to examine the stability of mutant proteins. After we generated clones for 204 known disease mutations using Clone-seq[12], we examined whether the mutant proteins could be stably expressed in human cells using the GFP assay. Compared with the corresponding wild-type proteins, the expression levels of 17 of the 204 (8.3%) mutants are significantly diminished (**Fig. 7a**). To validate these findings, we performed western blotting for 10 random mutants that are stably expressed and 10 random mutants with significantly diminished expression levels (**Fig. 7b**). All western blotting results agree perfectly with our GFP readings[12].

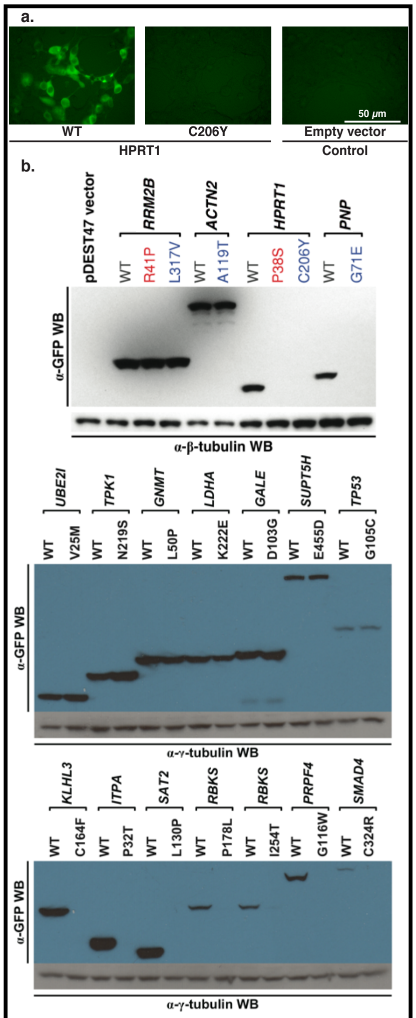


Fig 7. (a) GFP images of WT HPRT1 and its unstable mutant allele. (b) Western blot confirming results of our GFP assay

c.2.1.3. Four orthogonal high-throughput high-quality interaction-detection assays in our InPOINT pipeline. Current high-throughput interaction-detection technologies can benefit from an increase in sensitivity[18-20]. To address this, we have developed a high-throughput interaction-detection tool-kit[18,20,21] consisting of four complementary high-quality assays: Protein Complementation Assay (PCA)[22], yeast two-hybrid (Y2H), LUminescence-based Mammalian IntERactome mapping (LUMIER)[23], and 96-well-plate-based Nucleic Acid Programmable Protein Array (wNAPPA)[24]. With a large set of positive and negative controls for human proteins, we found that all four assays are of high quality and combining four assays significantly improves both sensitivity and specificity in detecting true protein interactions[19].

PRELIM
RES
?

c.2.1.4. Using our high-throughput InPOINT pipeline to examine the effects of disease mutations on protein interactions. We investigated whether these 204 mutations could affect protein-protein interactions using the four assays in our InPOINT pipeline. We found that 21 of the 27 (78%) "interface residue" mutations, 57 of the 100 (57%) "interface domain" mutations, and only 22 of the 77 (29%) "away from the interface" mutations disrupt the corresponding interactions, confirming that structural information of interactions greatly improves our understanding of the impact of disease mutations[12]. Y2H has been applied by us and other groups to examine hundreds of disease mutations and has been proven to be an effective approach[10-13,25]. The novelty of our InPOINT pipeline is that it combines four orthogonal assays (PCA, Y2H, LUMIER, and wNAPPA). Combining four orthogonal assays and using only consistent results by two or more assays will ensure scientific rigor and practically eliminate false-positives in our results.

c.2.2. Research Design

c.2.2.1. High-throughput cloning of ~500 WT TREs and ~2500 non-coding SNPs on these TREs using Clone-seq. Sequence-specific forward and reverse primers containing attB1 and attB2 sequences for 769 WT TREs will be designed by our automated online primer design website "http://primer.yulab.org"[12], and synthesized in bulk as "Trumer Oligo" plates by Eurofins Genomics. Using human genomic DNA as template, the selected TREs will be PCR amplified in 96-well format with high-fidelity Phusion DNA polymerase to minimize introduction of unintended mutations. 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[12]. Briefly, *E. coli* transformation is performed and a 20 μL aliquot of the cells is then spread onto LB + Spectinomycin plates in high-throughput using the Tecan robot. The next day, four colonies per allele are picked for Illumina sequencing using QPix-HT. After identifying successful clones without any unwanted mutations through our customized variant calling pipeline, we robotically picked out these 769 WT TRE clones for downstream experiments.

Primers for site-directed mutagenesis are designed by our automated online primer design website "http://primer.yulab.org"[12], and synthesized in bulk as "Trumer Oligo" plates by Eurofins Genomics. The mutant clones will be generated using our Clone-seq protocol[12]. Briefly, 50 μL mutagenesis PCR reactions are set up on ice in 96-well PCR plates using Phusion polymerase (NEB M0530) according to manufacturer's manual with WT TRE clones generated above. PCR products are digested by *Dpn*I (NEB R0176L) overnight at 37 °C. *E. coli* transformation, colony picking, and Illumina sequencing will be performed as described above through our high-throughput protocol using Tecan and QPix-HT robots. After identifying successful clones with the designed SNP but without any unwanted mutations through our customized variant calling pipeline, we robotically pick out the 1,407 successful mutant TRE clones for downstream experiments.

These fully sequence-verified WT and mutant entry clones will be subjected to Gateway LR reaction to transfer TREs in the entry vector to our modified pDEST-iSTARR destination vectors via recombination. The resulting expression clones will be pooled, maxipreped, and subjected to iSTARR-seq analysis in colon organoids.

c.2.2.2. Quantitatively measuring enhancer activity of WT and mutant TREs using iSTARR-seq. The 1,407 SNPs and their corresponding WT entry clones generated in **c.1.2.1** will be cloned into both pDEST-iSTARR-dCP and pDEST-iSTARR-hkCP vectors by Gateway LR reaction. In order to produce lentiviral particles carrying an iSTARR-seq library, the iSTARR-seq library plasmids will be transfected into HEK293T cells together with the envelope plasmid and the packaging plasmids. The viral particles will be collected from the culture medium of the transfected cells at 60h post transfection and then titrated with qRT-PCR targeting the viral RNA. Colon organoids will be transduced with the harvested lentiviral particles at desired MOI and selected with puromycin. Towards the end of the selection process, the integration rate will be confirmed by qPCR with genomic DNA (gDNA) extracted from a small portion of the transduced cells. The cells will then be collected for gDNA and total RNA extraction. mRNA derived from iSTARR-seq vectors will first be reverse transcribed and then PCR-amplified according to previous publication[14] with minor modifications. Briefly 1st-strand cDNA will be synthesized by reverse transcription with a vector backbone-specific primer annealing to 3'-end of the transcripts. Each primer molecule will contain a unique 15 nt molecular barcode to label each cDNA molecule (**Fig. 2**). Two rounds of nested PCR with low cycle numbers will be performed to amplify the TRE region in the cDNA without introducing contamination from transfected plasmid DNA or copy number bias. The cDNA library will be subjected to tagmentation with Tn5 transposase and customized sequencing adaptors containing indexing barcodes. After another round of low-cycle PCR for enriching successfully tagmented cDNA fragments, the barcoded library will then be sequenced with Illumina HiSeq or NextSeq.

Another sequencing library targeting gDNA-integrated TREs in the transduced cells will also be prepared and sequenced using the similar procedure as that for the mRNA. In addition, the lentiviral library will also be processed and sequenced as a control for overall library quality. The total number of the mRNA or DNA molecules of a given TRE (WT and all the mutants) will be the number of unique molecular barcodes associated with it. The proportion of each mutant is calculated based on the number of sequencing reads at its corresponding mutation site. The transactivity of a specific allele of a TRE (WT or mutant) will be calculated as the ratio of the number of mRNA molecules derived from the allele over the number of the TRE allele integrated into the gDNA.

c.2.2.3. High-throughput dual luciferase assays to further confirm and nominate functional non-coding risk variants. The canonical 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. All WT and mutant TREs 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 normal colon organoid cells by electroporation. The activity of each of the WT and mutant TREs as indicated by the intensity of bioluminescence will be measured by with Dual-Glo luciferase assay system (Promega).

c.2.2.4. High-throughput site-directed mutagenesis to generate ~1500 coding mutants through Clone-seq. Clone-seq will be carried out as described in our previous publication[12] and **c.1.2.1**. All WT clones are obtained from the Human ORFeome 8.1[26], which is a fully sequence-verified Gateway-compatible ORF clone library for human genes that we have purchased and maintained for the past five years. After Illumina sequencing, correct clones without any unwanted mutations are identified using our customized variant calling software[12].

c.2.2.5. High-throughput InPOINT pipeline (GFP assay) to test the stability of the ~1500 mutant proteins. All WT and mutant clones are first moved into the pDEST-GFP-mCherry vector using automated Gateway LR reactions in 96-well format. A 100 ng aliquot of the expression clone is used for transfection into HEK293T cells in 96-well plates using polyethylenimine. At approximately 48 hrs post-transfection, fluorescence intensities of transfected cells are measured with a Tecan M1000 at 395/507 nm for cycle 3 GFP (Invitrogen) and 580/612 nm for mCherry, denoted as I_g and I_r, respectively. As negative controls, the GFP and mCherry fluorescence intensities corresponding to cells transfected with the empty pDEST-GFP-mCherry vector (with a plate-specific mean I_{gb} and s.d. σ_{gb}) and empty pcDNA-DEST47 vector (with a plate-specific mean I_{rb} and s.d. σ_{rb}) are measured. A plate-specific Z_g and Z_r are calculated as $Z_g = (I_g - I_{gb})/\sigma_{gb}$ and $Z_r = (I_r - I_{rb})/\sigma_{rb}$. A WT clone is considered to have stable expression if its Z_g and Z_r values are both > K. Here, K = 1.645, corresponding to the single tail P value of 0.05 for a normal distribution (i.e., it has significantly higher expression than background for both GFP and mCherry). For mutants with corresponding stable WTs, we remove transfection failures ($Z_r \leq K$) and then calculate normalized ***quantitative*** stability scores for both WT and mutant:

$$S_{WT} = \left(\frac{I_g - I_{gb}}{I_r - I_{rb}}\right)_{WT} \quad and \quad S_{mut} = \left(\frac{I_g - I_{gb}}{I_r - I_{rb}}\right)_{mut}$$

All experiments will be performed in triplicate. Mutations that significantly affect protein stability will be identified by comparing the means of log(S_{WT}) and log(S_{mut}) scores using a ***t***-test (the log transformed stability scores follow a normal-like distribution). We will calculate a ***quantitative*** relative stability index, $RSI = \overline{S_{mut}}/\overline{S_{WT}}$, for mutations that significantly affect protein stability. To further ensure the quality of our results, we will perform an ELISA assay using anti-FLAG antibody for all 121 mutants. This is part of the LUMIER assay that we routinely apply to test the presence of the bait protein. Only mutants with consistent results between GFP and ELISA assays will be kept for downstream analyses, ensuring data quality and scientific rigor.

c.2.2.6. High-throughput InPOINT pipeline to test the effects on interactions of the ~1500 mutant proteins. Next, we will examine the impact of mutations on specific interactions: (1) **PCA.** Briefly, mutant ORF clones will be transferred by Gateway LR reactions into vectors encoding the two fragments of YFP (Venus variant) fused to the N-terminus of the tested proteins. Baits were fused to the F1 fragment (amino acids 1-158 of YFP) and preys to the F2 fragment (amino acids 159-239 of YFP). Plasmids encoding the two proteins are used for transfection into HEK293T cells in 96-well plates, using Lipofectamine2000 (Invitrogen). 48 hrs post-transfection cells are processed with Tecan M1000. A pair are considered interacting if the YFP fluorescence intensity is ≥2 fold higher over background. (2) **LUMIER.** ORFs are cloned into Gateway-compatible LUMIER vectors by LR reactions and minipreped. HEK293T cells were transfected in 96-well plates. After transfection, cells are processed for immunoprecipitation. LUMIER intensity ratio (LIR) values are obtained for the immunoprecipitates (LIR-IP) and calculated similarly for the total lysates (LIR-TOT). Normalized LIR (NLIR) was calculated as the ratio LIR-IP/LIR-TOT. A pair with NLIR score of ≥ 33.2 are considered to be interacting. (3) **Y2H.** ORFs are cloned into pDEST-AD and pDEST-DB vectors by LR reactions. All DB-X and AD-Y plasmids will be transformed individually into the Y2H strains *MAT*α Y8930 and *MAT***a** Y8800, respectively. After mating, only yeast cells containing interacting pairs of DB-X and AD-Y will grow on selective media (i.e., expression of *HIS3* and *ADE2* reporter genes). (4) **wNAPPA.** ORFs are cloned into pCITE-HA and pCITE-GST vectors by LR reactions. Both prey and bait plasmids are added to Promega TnT coupled transcription-translation mix and incubated to express proteins. The whole mix is then added to anti-GST antibody-coated 96-well plates. After binding and capture, plates are incubated with primary and secondary antibody and visualized using chemiluminescence with Tecan M1000. Wells with ≥3 fold higher intensity over background in either configuration are scored positives. Only disruptions confirmed by two or more assays (including Y2H) will be considered disrupted for all downstream analyses. **Combining four orthogonal assays and using only consistent results by two or more assays will ensure the quality and practically eliminate false-positives in our results, ensuring scientific rigor.**

C.3. AIM 3 Medium-throughput *in vivo* quantification of cellular phenotypes and validation of 10 coding and non-coding variants in prostate organoids

In aim 3, we will take variants prioritized based on molecular impact and investigate their effects on cellular- and organoid-level phenotypes related to cancer.

C.3.A. Preliminary results

C.3.A.1 Experience with CRISPR-Cas9 and interaction with P54 Cancer Systems Biology Center. In our work, we will take advantage of collaborative interactions with the NCI-funded P54 Cancer Systems Biology Center at Yale (CaSB@Yale), directed by Prof. Andre Levcheno (see his letter attached). In particular, we will use the service of its Core 2 focused on the CRISPR-based generation of cell and animal cancer models, involving knockin and knockout of pre-determined molecular target. Previously and currently, researchers at the Center have applied *in vivo* somatic genome editing to generate tumor models of specific driver genes in cell lines and mouse models of different cancers (see Fig. 1 for an example of liver cancer).

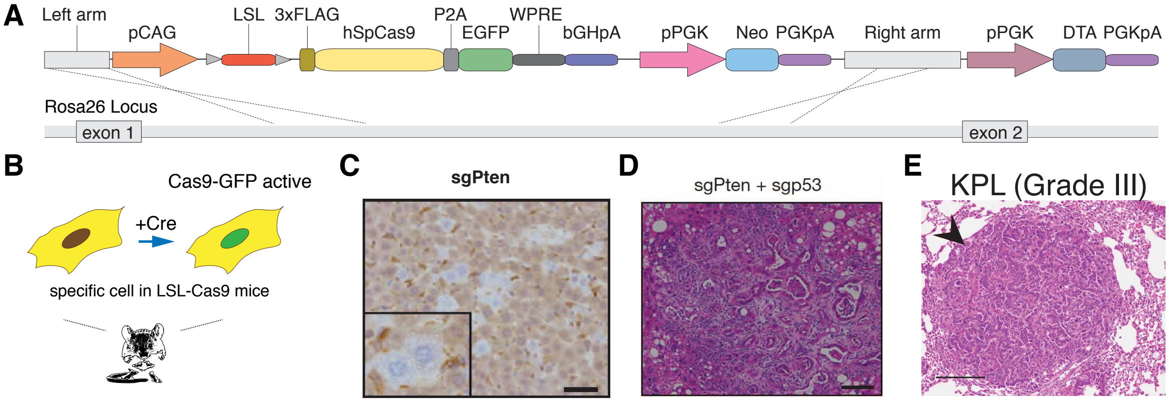


Fig. 1 Demonstration of in vivo genome editing for cancer modeling. (A) Schematics of a conditional Cas9 knockin mice; (B) Schematics of Cas9 activation in tissues of interest upon delivery of Cre; (C) Example of single gene targeting by CRISPR, hydrodynamic delivery of Cas9 and sgRNA plasmid targeting Pten in mouse liver leads to clonal Pten null cells; (D) liver cancer model induced by hydrodynamic delivery of Cas9 and two sgRNAs targeting Pten and p53; (E) lung adenocarcinoma model using gene editing with conditional Cas9 knockin mice, showing istology of AAV-KPL generated lung cancer (grade III).

Cas9 was targeted to cells and animals to generate specific genetic changes that can promote oncogenesis or model other mutations (Xue et al., 2014). Viral delivery enables targeting of almost any tissue, including prostate tissues, constrained by the packaging capacity, which limits the number of sgRNAs, HDR donors, and other elements that will fit within the same vector as Cas9. For example, lentivirus mediated delivery was used to deliver Cas9 in mammalian cells to study cancer (Shalem et al., 2014, Wang et al. 2014, Chen et al. 2015). Adeno-associated viral (AAV) vectors are DNA-based and not prone to recombination, making the expression of multiple U6-sgRNA cassettes feasible. CaSB@Yale generated a Cre-conditional Cas9 mouse model, which facilitates rapid and efficient modeling of single and multi-genic mutations in specific tissue and cell types of interest. In this model, Cas9 is already present and dormant within the genome of all cells, which opens up a larger capacity for delivery of sgRNAs as well as other elements. We have combined this conditional Cas9 mouse with AAV vector-mediated expression of sgRNAs in the lung, and modeled lung cancer using a combination of the *Kras* oncogene and two tumor suppressor genes, *p53* and *Lkb1* (*Stk11*) (Platt et al. 2014). These enable novel viral vector based platforms to study the combinatorial contribution of mutations, defining tumor phenotypes and their evolution in vitro and in vivo.

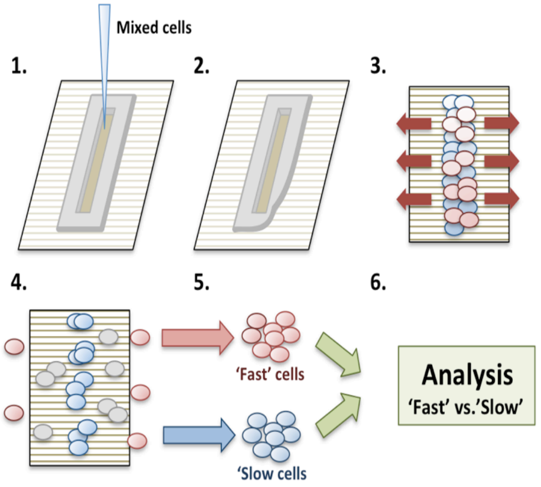


Figure 2. The steps involved in the RACE based phenotypic filtering phenotypic filtering into relatively highly proliferative and highly migratory cells

C.3.A.3 Assays for cell proliferation and migration

The Director of CaSB@Yale, Andre Levchenko, will also participate in the proposed work by contributing the analysis platforms allowing to separate prostate cancer populations into sub-populations of highly migratory and highly proliferative cells. Co-existence of such sub-populations is common, stemming from what is commonly referred to as the 'go or grow' phenotypic switch, which can in turn be controlled by various genetic and environmental alterations. In our prior work, we have demonstrated that highly migratory and highly proliferative cells can be separated using anisotropically nanofabricated substrata (ANFS) that closely mimic the fibrous structure of ECM [18-20]. This ANFS, in addition to mimicking putative alignment of ECM fibers [21], converts cell migration from a 2D random walk to an essentially 1D persistent and unidirectional movement along the direction of nano-fibers – similar to cell migration and alignment observed in sparse 3D ECM (Fig 2A). This similarity of migration on ANFS to 3D cell migration *in vivo* not only suggests that the analysis is biomimetic and more relevant than the usual Petri dish experimentation, but also provides a convenient way to contrast the migration of differentially perturbed cells against each other. **In this proposal, we will develop the initial screening of aggressive melanoma cells. We refer to this first test as the Rapid Analysis of Cell migration Enhancement (RACE).**

C.3.A.3. Patient-derived tumor organoids as a tool for precision cancer care. We recently demonstrated that we can develop cancer and benign organoids. From a cohort of 145 specimens from patients with advanced cancers including prostate (52). We were able to develop tumor organoids from 38.6%. We define successful establishment of PDTO cultures when they contain viable cells that form spheroid-like structures and can be propagated after the initial processing for at least five passages. These specimens are characterized, stored in our living biobank and are used for functional studies. Cell viability was assessed in the first ten established cultures at passages 2-4, and in 9 out of 10 cases, > 90% of cells were viable. Tumor and benign organoids are characterized using cytology and histology as previously described [21]. As the data is now published we only note that we have been able to perform extensive studies with these organoids including CRIPSR-cas9 manipulation (FANCA PAPER), drug screens (PAULI REFERENCE), and lenti-viral SH infection. With many years experience, we are confident that developing benign prostate cell lines for this Aims should be readily accomplished.

C.3.B. Research Plan.

C.3.B.1. Medium throughput quantification of variant effects on cell proliferation and migration. We will introduce knockin-based perturbations of the top 120 highly scored genetic targets using the CRISPR techiques outlined above and study their effect on cell migration and proliferation using RACE. In our preliminary analysis, as a proof-of –principle, we employed the virally packaged, Doxycycline-inducible short hairpin RNAs (shRNAs) collection. The key aspect of this library, is that the silencing is inducible, and the constructs are barcoded. The cells have been transfected by a pool of shRNAs en-mass and scored for a phenotype of choice, e.g., cell migration and prolfieration. Cells achieving a high or low score can then be assayed for the presence of specific shRNAs by PCR-amplifying it using a set of unique primers. As described in Fig 2, the mixed populations of melanoma cells transduced with different shRNAs are plated on ANFS in a small slit of a poly-(di)-methyl-syloxene (PDMS) stencil. After cell attachment, the stencil is removed, allowing the cells to migrate along the direction of nano-ridges. The silencing of expression of sub-sets of specific genes by the corresponding shRNAs differentially affects cell movement or proliferation, so that the RACE results in cells 'racing' with different values of speed and persistence along the direction specified by the ANFS. After 1 week of RACE, we harvest the 1/3 of the ANFS area where cells were originally seeded ('slow group', high prolifeation) and the most distant 1/3 from the area of cell seeding ('fast group'; high migration). The cells from the 'fast' and 'slow' groups are re-seeded separately and the RACE assay is repeated three times to enrich the 'fast' and 'slow' populations through sequential 'racing' periods. Finally, the cells were harvested, and the shRNA sequences were PCR-amplified to determine the genes whose silencing affected the phenotype. **<u>As a result, we can determine the role of silenced genes in driving cell migration/proliferation and thus potentially the propensity of cancer cells to initiate invasion and metastasis.</u>** Importantly, we have shown that this assay selects cells not only for increased migration but also for other characteristics (signaling, stem-ness, metabolic control, etc.) (a sample of the results in a melanoma cell line is shown in Fig 3).

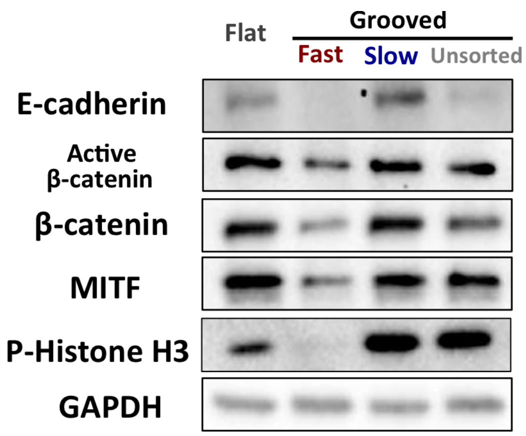


Figure 3. Partial characterization of the 'fast' and 'slow' cell sub-populations with the cell population of SK-Mel-28 melanoma cell line.

C.3.B.2 Validation of 10 coding and non-coding variants in prostate organoids

C.3.B.2.a. Specimen procurement. Patient-derived fresh tissue samples will be collected with written informed patient consent in accordance with the Declaration of Helsinki and with the approval of the Ethics Board at the University of Bern and the Inselspital (Bern Hospital Group). Fresh tissue biopsies and resection specimens are taken directly in the procedure rooms. Fresh tissue biopsies will be transported to the laboratory to establish primary tumor organoid cultures. Macroscopically different appearing tumor areas will be collected and processed individually. The time between harvesting fresh tissue specimens and placing them in transport media [Dulbecco's modified Eagle medium (DMEM, Invitrogen) with Glutamax (1x, Invitrogen), 100U/ml penicillin, 100ug/ml streptomycin (Gibco), Primocin 100ug/ml (InvivoGen), 10 uM Rock inhibitor Y-27632 (Selleck Chemical Inc)] should be less than 30 minutes.

C.3.B.2.b. Tissue processing and cell culture conditions. Tissue samples will be washed a minimum of three times with transport media and placed in a sterile 3 cm petri dish (Falcon) for either total mechanical dissociation or dissection into smaller pieces (~2 mm diameter) prior to enzymatic digestion. Enzymatic digestion was done with 2/3rd of 250 U/mL collagenase IV (Life Technologies) in combination with 1/3rd of 0.05% Trypsin-EDTA (Invitrogen) in a volume of at least 20 times the tissue volume. The cells will be resuspended in a small volume of tissue-type specific primary culture media with a 1:2 volume of growth factor reduced Matrigel (Corning).

CRISPR-cas9 Experiments. We will employ CRISPR-cas9 gene exiting approaches as described above to modify benign luminal prostate organoids. Analysis with regards to downstream effects will be compared to scrambled guide RNA treated cell lines. (**QUESTIONS FOR GROUP CAN WE DEFINE OUR READOUTS HERE)**

YU REFERENCES CITED:

1. Vidal, M.*, et al.* Interactome networks and human disease. ***Cell*** 144, 986-998 (2011).
2. Barabasi, A.L.*, et al.* Network medicine: a network-based approach to human disease. ***Nat Rev Genet*** 12, 56-68 (2011).
3. Pawson, T. & Nash, P. Protein-protein interactions define specificity in signal transduction. ***Genes Dev*** 14, 1027-1047 (2000).
4. Cruchaga, C.*, et al.* Rare coding variants in the phospholipase D3 gene confer risk for Alzheimer's disease. ***Nature*** 505, 550-554 (2014).
5. MacArthur, D.G.*, et al.* A systematic survey of loss-of-function variants in human protein-coding genes. ***Science*** 335, 823-828 (2012).
6. Cox, A.*, et al.* A common coding variant in CASP8 is associated with breast cancer risk. ***Nat Genet*** 39, 352-358 (2007).
7. 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).
8. Khurana, E.*, et al.* Integrative annotation of variants from 1092 humans: application to cancer genomics. ***Science*** 342, 1235587 (2013).
9. Guo, Y.*, et al.* Dissecting disease inheritance modes in a three-dimensional protein network challenges the "guilt-by-association" principle. ***Am J Hum Genet*** 93, 78-89 (2013).
10. Wang, X.*, et al.* Three-dimensional reconstruction of protein networks provides insight into human genetic disease. ***Nat Biotechnol*** 30, 159-164 (2012).
11. Sahni, N.*, et al.* Widespread macromolecular interaction perturbations in human genetic disorders. ***Cell*** 161, 647-660 (2015).
12. Wei, X.*, et al.* A massively parallel pipeline to clone DNA variants and examine molecular phenotypes of human disease mutations. ***PLoS Genet*** 10, e1004819 (2014).
13. Zhong, Q.*, et al.* Edgetic perturbation models of human inherited disorders. ***Mol Syst Biol*** 5, 321 (2009).
14. Arnold, C.D.*, et al.* Genome-wide quantitative enhancer activity maps identified by STARR-seq. ***Science*** 339, 1074-1077 (2013).
15. Schwartz, J.J.*, et al.* Accurate gene synthesis with tag-directed retrieval of sequence-verified DNA molecules. ***Nat Methods*** 9, 913-915 (2012).

16. McKenna, A., *et al.* The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. ***Genome Res*** 20, 1297-1303 (2010).
17. Zabidi, M.A., *et al.* Enhancer-core-promoter specificity separates developmental and housekeeping gene regulation. ***Nature*** 518, 556-559 (2015).
18. Braun, P., *et al.* An experimentally derived confidence score for binary protein-protein interactions. ***Nature methods*** 6, 91-97 (2009).
19. Venkatesan, K., *et al.* An empirical framework for binary interactome mapping. ***Nat Methods*** 6, 83-90 (2009).
20. Yu, H., *et al.* High-quality binary protein interaction map of the yeast interactome network. ***Science*** 322, 104-110 (2008).
21. Yu, H., *et al.* Next-generation sequencing to generate interactome datasets. ***Nat Methods*** 8, 478-480 (2011).
22. Remy, I. & Michnick, S.W. Mapping biochemical networks with protein-fragment complementation assays. ***Methods Mol Biol*** 261, 411-426 (2004).
23. Barrios-Rodiles, M., *et al.* High-throughput mapping of a dynamic signaling network in mammalian cells. ***Science*** 307, 1621-1625 (2005).
24. Ramachandran, N., *et al.* Next-generation high-density self-assembling functional protein arrays. ***Nat. Methods*** 5, 535-538 (2008).
25. Fuxman Bass, J.I., *et al.* Human gene-centered transcription factor networks for enhancers and disease variants. ***Cell*** 161, 661-673 (2015).
26. Yang, X., *et al.* A public genome-scale lentiviral expression library of human ORFs. ***Nat Methods*** 8, 659-661 (2011).

RUBIN References

1 Barretina, J. *et al.* The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature* **483**, 603-607, doi:10.1038/nature11003 (2012).

2 Shoemaker, R. H. The NCI60 human tumour cell line anticancer drug screen. *Nat Rev Cancer* **6**, 813-823, doi:10.1038/nrc1951 (2006).

3 Iorio, F. *et al.* A Landscape of Pharmacogenomic Interactions in Cancer. *Cell* **166**, 740-754, doi:10.1016/j.cell.2016.06.017 (2016).

4 Iorio, F. *et al.* A Landscape of Pharmacogenomic Interactions in Cancer. *Cell*, doi:10.1016/j.cell.2016.06.017 (2016).

5 Paul, S. M. *et al.* How to improve R&D productivity: the pharmaceutical industry's grand challenge. *Nat Rev Drug Discov* **9**, 203-214, doi:10.1038/nrd3078 (2010).

6 Li, A. *et al.* Genomic changes and gene expression profiles reveal that established glioma cell lines are poorly representative of primary human gliomas. *Mol Cancer Res* **6**, 21-30, doi:10.1158/1541-7786.MCR-07-0280 (2008).

7 Baker, L. A., Tiriac, H., Clevers, H. & Tuveson, D. A. Modeling pancreatic cancer with organoids. *Trends Cancer* **2**, 176-190, doi:10.1016/j.trecan.2016.03.004 (2016).

8 Boj, S. F. *et al.* Organoid models of human and mouse ductal pancreatic cancer. *Cell* **160**, 324-338, doi:10.1016/j.cell.2014.12.021 (2015).

9 Clevers, H. Modeling Development and Disease with Organoids. *Cell* **165**, 1586-1597, doi:10.1016/j.cell.2016.05.082 (2016).

10 Gao, D. *et al.* Organoid cultures derived from patients with advanced prostate cancer. *Cell* **159**, 176-187, doi:10.1016/j.cell.2014.08.016 (2014).

11 Huang, L. *et al.* Ductal pancreatic cancer modeling and drug screening using human pluripotent stem cell- and patient-derived tumor organoids. *Nat Med* **21**, 1364-1371, doi:10.1038/nm.3973 (2015).

12 Nash, C. E. *et al.* Development and characterisation of a 3D multi-cellular in vitro model of normal human breast: a tool for cancer initiation studies. *Oncotarget* **6**, 13731-13741, doi:10.18632/oncotarget.3803 (2015).

13 Baker, B. M. & Chen, C. S. Deconstructing the third dimension: how 3D culture microenvironments alter cellular cues. *J Cell Sci* **125**, 3015-3024, doi:10.1242/jcs.079509 (2012).

14 Jamieson, L. E., Harrison, D. J. & Campbell, C. J. Chemical analysis of multicellular tumour spheroids. *Analyst* **140**, 3910-3920, doi:10.1039/c5an00524h (2015).

15 Pampaloni, F., Reynaud, E. G. & Stelzer, E. H. The third dimension bridges the gap between cell culture and live tissue. *Nat Rev Mol Cell Biol* **8**, 839-845, doi:10.1038/nrm2236 (2007).

16 Barbone, D., Yang, T. M., Morgan, J. R., Gaudino, G. & Broaddus, V. C. Mammalian target of rapamycin contributes to the acquired apoptotic resistance of human mesothelioma multicellular spheroids. *J Biol Chem* **283**, 13021-13030, doi:10.1074/jbc.M709698200 (2008).

17 Frankel, A., Man, S., Elliott, P., Adams, J. & Kerbel, R. S. Lack of multicellular drug resistance observed in human ovarian and prostate carcinoma treated with the proteasome inhibitor PS-341. *Clin Cancer Res* **6**, 3719-3728 (2000).

18 Mueller-Klieser, W. Three-dimensional cell cultures: from molecular mechanisms to clinical applications. *Am J Physiol* **273**, C1109-1123 (1997).

19 Mueller-Klieser, W. Tumor biology and experimental therapeutics. *Crit Rev Oncol Hematol* **36**, 123-139 (2000).

20 Pickl, M. & Ries, C. H. Comparison of 3D and 2D tumor models reveals enhanced HER2 activation in 3D associated with an increased response to trastuzumab. *Oncogene* **28**, 461-468, doi:10.1038/onc.2008.394 (2009).

21 Pauli, C. *et al.* An emerging role for cytopathology in precision oncology. *Cancer Cytopathol* **124**, 167-173, doi:10.1002/cncy.21647 (2016).

Page 7: [1] Deleted Michael Rutenberg Schoenberg 6/11/17 1:04:00 PM

FunSeq and FunSeq2 allow us to score mutations based on predicted molecular functional impact. Variants with high FunSeq scores are predicted to be functionally impactful variants. These high scoring variants tend to be located in functionally significant noncoding domains, and may correspond to undiscovered drivers (both strong & weak) as well as passenger variants that decrease tumor cell fitness. Conversely, common variations tend to arise in functionally unimportant regions due to constraint by selective pressure. Thus, genomic features that are enriched with common polymorphisms are less likely to contribute to the deleteriousness of variants and so receive lower scores.

We will expand the scoring system of FunSeq\cite{24092746} and Funseq2\cite{25273974} in order to integrate the additional variant attributes we measure. In general, features can be classified as discrete (e.g., either within or outside of a given functional annotation) or continuous (e.g., the PWM change in ‘motif-breaking’). We will weigh these two sets of features using different strategies. For each discrete feature, we will calculate the probability that it overlaps with common polymorphisms. We will then calculate the information content to denote the value of discrete features. 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 , which is associated with a value , the probability is first estimated using common variants: . The score of continuous feature is defined as .**

The score () is calculated as . We will also incorporate the feature dependency structure when calculating the scores by removing redundant features using feature selection or by performing dimensionality reduction