A combined pipeline for prioritizing prostate cancer variants, integrating computational prioritization with molecular, cellular, and organoid-level phenotypes

Systems biology U01 grant
June 4, 2017

Word counts:
Specific aims: 1,026
Significance: 923
Aim 1: 2,540
Aim 2: 2,617
Aim 3 (Rubin component): 826
Total (excluding references): 7,932

SPECIFIC AIMS

In this proposal, briefly, we plan to develop mathematical models to prioritize and rank non-coding and coding mutations in similar terms. These models will rank the impact of mutations causing cancer in terms of their underlying genomic alteration. We will then assay the actual phenotypes produced by these mutations on three scales: molecular activity, cellular phenotypes, and phenotypes in cultured organoids. Doing these experiments will produce a data resource of prioritized mutations and iterated mathematical models for prioritizing them as a product. It will also allow us to address a number of questions about cancer.

First of all, cancer genomics has revealed that there are often thousands of mutations per tumor genome but only a small fraction of them are in coding regions. Yet, almost all of the known driver mutations in cancer are in coding regions. Is this because, fundamentally, non-coding mutations have less impact than coding ones, or just simply because of an ascertainment bias on our part?

Second of all, is it the case that a mutation prioritized to give a strong impact in terms of effect on molecular networks binding will also have a strong effect on cellular phenotype and this will have also a strong effect on organismal phenotypes such as contracting cancer. It's not clear that we'll see a similarity between these three levels and we will be able to ascertain that here.

To focus our analysis, we will prioritize both coding and noncoding variants in linked enhancers and promoters on a matched set of genes, including both validated and putative cancer drivers, as well as some control genes with no known cancer association. Non-coding mutations are potentially directly involved in our regulatory networks sitting in regulatory regions of the genome and they can be matched, in a system sense, to many of the coding mutations which directly effect protein-protein interfaces involved in protein networks. One question we will investigate is 'Are these mutations in any sense comparable or are, fundamentally, the coding mutations more deleterious?'

AIM 1 Computational prioritization of coding and non-coding somatic mutations

First, we will do this in a classical sense by looking for mutations under positive selection in cohorts that are recurrent in particular regions of the genome i.e. in particular domains of a protein or in particular non-coding elements and to do this we will use the recently constructed large datasets, e.g. from TCGA and PCAWG consortia. We will also prioritize mutations computationally by looking at their sequence level molecular impact. This will be done from using a variety of metrics such as: the degree to which the mutation directly breaks the functional site i.e. breaks the TF motif or protein-protein bind interface; the degree to which it effects central positions in the overall network; the degree to which it's associated with a site that has an obvious allelic effect and sensitivity to sequence; the degree to which it sits in a functional element; and the degree to which it shows obvious conservation across organisms or within the human population, for instance as measured from GERP score.

From the combination of positive selection and functional impact, we will develop mathematical models to prioritize mutations and lists to prioritize mutations that we will then hand off to the validation components of the proposal. We will take the results each year from the validation components and use it to refine our models by a variety of simple iterate machine running tactics such as a Bayesian or online conjugate gradient updates.

AIM 2 High-throughput *in vitro* quantification of molecular phenotypes of ~2500 non-coding and ~1500 coding mutations We will select ~500 coding and ~1000 non-coding mutations and subject them to a number of high-throughput in vitro assays to look at their molecular readout. We will take advantage of our novel Clone-seq pipeline to generate these mutant clones in large-scale. As an integral part of the Clone-seq pipeline, each mutant clone will be fully sequence verified by next-

generation sequencing to ensure quality. Furthermore, we will assay the non-coding mutations using eSTARR-Seq and Promoter-seq the coding mutations to quantify their effect on enhancer and promoter activities. We will also assay the coding mutations using our high-throughput protein-protein interactome screening methodology described in our previous publications[8-11], *IN*tegrated *P*r*O*tein *IN*teractome per*T*urbation screening (InPOINT). This pipeline combines six different functional assays to examine experimentally the impact of hundreds of coding variants on protein stability and specific protein-protein interactions. From this we will be able to rank this pool of ~1500 variants in terms of their strongest molecular readouts.

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

In this aim we will look at cellular phenotypes associated with hundreds of mutations and then investigate the effects of a smaller number of mutations in organoids that are more realistic models tumor environments. We will evaluate ~150 coding and ~150 non-coding mutations in terms of their phenotypes for cell growth and also invasiveness, which is related to metastasis, using a variety of cell-based assays. The mutations will be introduced into CCD-18Co [[Replace with prostate!]] cells through CRISPR/Cas9 mutagenesis. We then will select the top 10 coding and non-coding mutations and evaluate them in a realistic tissue system – organoids derived from normal prostate samples. We will see if these mutations are actually associated with promoting cancer in this model system and then show the degree to which we can find non-coding mutations with as much functional impact as coding ones. We will further investigate the mechanisms through which mutations lead to cancer. For non- coding mutations, we will test alterations in transcript levels, H3K27Ac/H3K4me3 marks and transcription factor binding, comparing gene-edited and isogenic control prostate organoids. For coding mutations, we will perform co-IP, protein stability and selected functional assays in gene edited and isogenic control organoids. Throughout the process, we will feedback the results of each of the assays into our overall computational model and prioritization scheme developing a more accurate scheme. So with each year of the grant we will develop a more accurate model, eventually culminating near the end of the grant with a highly accurate model and a refined prioritization list.

SIGNIFICANCE

The complexity of genetic variation associated with cancer demands approaches that can assess the effects of different types of variants. A wealth of annotated data are available due to advances in sequencing technologies and efforts by consortia like ENCODE and 1000 Genomes, what engenders the need for comprehensive computational, mathematical, and experimental methods and analyses. Accordingly, we will leverage our experience and tools to prioritize variants *in silico*, *in vitro*, and *in vivo* (see Aims 1-3). These variant prioritization methods align with initiatives to develop the field of precision medicine and help scientists and medical practitioners understand the significance of unique combinations of genetic variants of each cancer patient.

While we know that driver variants usually alter genes that control cell growth and division, we still need to prioritize variants with respect to their deleteriousness, especially because certain variants in tumor cells may be the result rather than the cause of cancer. In addition, driver-passenger dichotomy might be simplistic. Recent studies *\cite{26456849} \cite{23388632}* suggest that some passenger mutations may have a weak effect on tumor cell fitness and may in turn promote or inhibit tumor growth. These mutations have been called “mini-drivers” or “deleterious passengers.” From a tumor fitness perspective, three categories can thus emerge: positively-selected driver variants, neutrally-selected passenger variants, and negatively-selected deleterious passenger variants. We think that studying the interplaying effects of both weak positive and negative selection variants may also reveal valuable insights into tumor growth patterns.

According to the philosophy of molecular reductionism, there is a functional impact associated with any positively or negatively selected variant. On the one hand, evidence of such impact is well-established for positively-selected variants promoting tumor growth. On the other hand, the impact of rapid accumulation of weak and deleterious passenger variants - which undergo negative selection - needs to be further studied as it could adversely alter tumor cell fitness \cite{23388632}. In fact, our previous studies suggest that an intermediate category of variants between high-impact putative driver variants and low impact neutral passengers exists and does have an observable functional impact that varies across cancer types and genome elements. We hypothesize that this intermediate functional category includes undiscovered drivers, deleterious passengers, or both. Further investigation should be done to better understand the functional ramifications of deleterious passengers and their effects on the fitness of cancer cells.

We will concentrate on somatic variants in cancer. Such alterations occur in protein coding and noncoding regions of the genome, and both coding and noncoding variants may vary in degree of impact on cancer development or protein formation and function. Historically, there has been a bias 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 noncoding regions may also be critical to a better understanding of cancer biology. Accordingly, we will consider a combination of coding and noncoding variants.

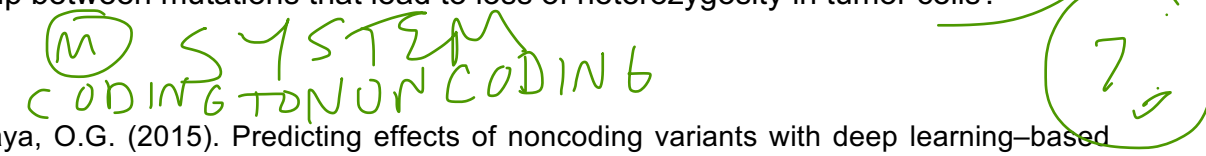
Effects of numerous genetic variants transcend the molecular level and propagate into the phenotype. However, the extents to which variant effects take place at the levels of molecular activity, cellular phenotype, and organismal phenotype are still unclear. The assumption that the impact of variants is consistent at all three levels needs to be examined. For that purpose, we plan to leverage our experience and use a variety of pipelines, cell-based assays, CRISPR-Cas9-based methods, and realistic prostate organoids. We will also study the relationship between different mutations and tumor growth and invasiveness (see Aims 2-3).

In our work, we will focus on prostate cancer. 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 the disease. Along with other major factors associated with prostate cancer such as the hormonal action of androgens and estrogens [2], more than 70 genetic susceptibility variants have been identified [3]. Suspected loci are continuously being discovered using GWAS studies [4] and genotyping arrays [5]. Such variants increase the predictability of the disease and have been associated with altering the expression levels of several genes.

Important genetic alterations associated with prostate cancer have effects on hormonal levels or take place in a variety of pathways. Among the known driver genes that prostate cancer shares with other cancer types — especially breast cancer being also a hormonal cancer — are tumor suppressing *BRCA1* and *BRCA2* [6]. Pathways targeted by DNA methylation in both prostate and breast cancers are also altered as a result of epigenetic modifications that depend on hormone receptor status and tumor recurrence. These modifications are associated with genes coding for zinc finger transcription factors and calcium binding proteins [7].

Other genetic variants detected across different types of prostate cancer take place in genes involved in lipid metabolism pathways. Among these genes are *MSMB, NUDT11*, *RBPMS2*, *NEFM*, and *KLHL33* [5]. Differential expression levels of genes active in focal adhesion, cell death, cell motility, and integrin signaling pathways have also been observed in the early stages of prostate cancer development [8]. Other important hormone-related alterations include the overexpression of *MYC*, *ERBB2* and *BCL2* genes [9].

In addition to prioritizing susceptible cancer variants, we will investigate the following related questions: (1) are noncoding variants as deleterious as coding ones *w.r.t.* prostate cancer incidence?, (2) do deleterious variants lead to the emergence of more deleterious ones in tumor cells?, (3) is there a fitness benefit for heterozygous *v.s.* homozygous mutation in tumor suppressor genes?, and (4) is there a relationship between mutations that lead to loss of heterozygosity in tumor cells?



References

[1] Zhou, J. & Troyanskaya, O.G. (2015). Predicting effects of noncoding variants with deep learning–based sequence model. Nat. Methods, 12, 931-943, PMCID: PMC4768299.

[2] Nelles, J. L., Hu, W. Y., and Prins, G. S. (2011). Estrogen action and prostate cancer. Expert Rev. Endocrinol. Metab., 6(3) 437-451, PMCID: PMC3134227.

[3] Eeles, R.A., Olama, A.A., Benlloch, S., Saunders, E.J., Leongamornlert, D.A., Tymrakiewicz, M. *et al.* (2013). Identification of 23 new prostate cancer susceptibility loci using the iCOGS custom genotyping array. Nat. Genet., 45, 385–91, PMCID: PMC3832790.

[4] Olama, A.A. *et al.* (2014). A meta-analysis of 87,040 individuals identifies 23 new susceptibility loci for prostate cancer. Nat. Genet., 46(10), 1103-9, PMCID: PMC4383163.

[5] Penny, K.L. (2015). Association of Prostate Cancer Risk Variants with Gene Expression in Normal and Tumor Tissue. Cancer Epidemiol. Biomarkers Prev., ;24(1), 255-60, PMCID: PMC4294966.

[6] Szulkin, R. (2015). Prediction of individual genetic risk to prostate cancer using a polygenic score. Prostate, 75(13), 1467-74, DOI: 10.1002/pros.23037.

[7] Day, T. K. and Bianco-Miotto, T. (2013). Common gene pathways and families altered by DNA methylation in breast and prostate cancers. Endocr Relat Cancer, 20(5), 15-32, DOI: 10.1530/ERC-13-0204.

[8] Gorlove, I.P. *et al.* (2009). Candidate pathways and genes for prostate cancer: a meta-analysis of gene expression data. BMC Med Genomics, 2, PMCID: PMC2731785.

[9] Mazaris, E. and Tsiotras, A. (2013). Molecular Pathways in Prostate Cancer. Nephrourol. Mon., 5(3), 792–800, PMCID: PMC3830904.

AIM 1 Computational prioritization of coding and non-coding somatic mutations

In aim 1, we will prioritize both coding and noncoding colon cancer variants for investigation assays of molecular, cellular, and organoid-level phenotypes, simultaneously validating candidate oncogenic variants and refining tools to predict impactful variants (Figure 1). In doing this we will leverage our extensive experience in both variant prioritization and cancer genome analysis. We have developed numerous tools for both coding and noncoding variants, using a variety of approaches.

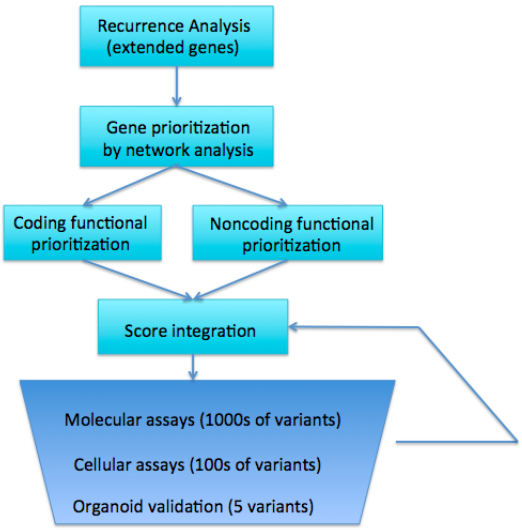


Figure 1. Overall variant prioritization workflow

A. Prior Experience

Experience prioritizing protein-coding variants

We have developed a number of tools that search for deleterious protein-coding variants. Since minor disruptions to some protein-coding genes can cause disease, while other genes can experience total loss of function with no observable effect, it is important to identify which genes have important functions that could cause disease when altered. Our netSNP tool integrates protein-protein, transcription factor, and metabolic networks to build a classifier that distinguishes genes that essential from those that are loss of function tolerant (Fig 1) \cite{23505346}. To analyze specific mutations within genes, our variant annotation tool is a utility that helps identify variants that overlap genes or other annotations, including, for example, whether variants induce premature stop codons\cite {22743228}. Building upon this, we have developed a pipeline for Analysis of Loss of Function Transcripts (ALoFT) that predicts whether mutations will cause loss of function in genes, and whether loss of one copy or both copies of a given gene is sufficient to cause disease and have applied this tool to cancer genomes, showing an enrichment for predicted loss of function mutations in known cancer-associated genes. Beyond identification of genes whose mutation can cause disease, we have also developed tools that characterize the effects of specific variants. Our STRESS tool identifies mutations that might affect allosteric hotspots in proteins, which can be key to protein function \cite{27066750}. Along similar lines, our Frustration tool uses calculations of localized structural frustration to identify key functional protein regions \cite{27915290}. Finally, our Intensification tool searches for deleterious mutations particularly within repeat regions of proteins \cite{27939289}.

Experience in noncoding genome analysis and allelic analysis

Our interest and expertise in prioritizing noncoding DNA variants rests on our experience analyzing a wide array of genomic assays to characterize noncoding genomic elements. 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}, identify and functionally categorize new noncoding transcripts \cite{21177971, 25164757}, and to predict enhancer regions \cite{22950945}, including some that have been functionally validated \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, which show chromosome conformations that can aid enhancer-target linkage inference \cite{28369339, http://biorxiv.org/content/early/2016/12/29/097345}. In addition to identifying, quantifying, and linking noncoding noncoding genomic elements, we have been multiple linear and nonlinear models use epigenetic signals to predict gene expression \cite{22955978, insert others}. Moreover, we have extensive experience building genomics data into networks that help explain gene regulation and to identify key regulators \cite{22955619}.

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 from the gEUVADIS \cite{24037378} and ENCODE \cite{22955616}. 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 genome elements, indicating that these elements are particularly sensitive to mutations.

Experience in noncoding variant prioritization

We have extensively analyzed patterns of variation in noncoding regions, along with their coding targets[90,95,114]. In recent studies[26,27], 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. 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.

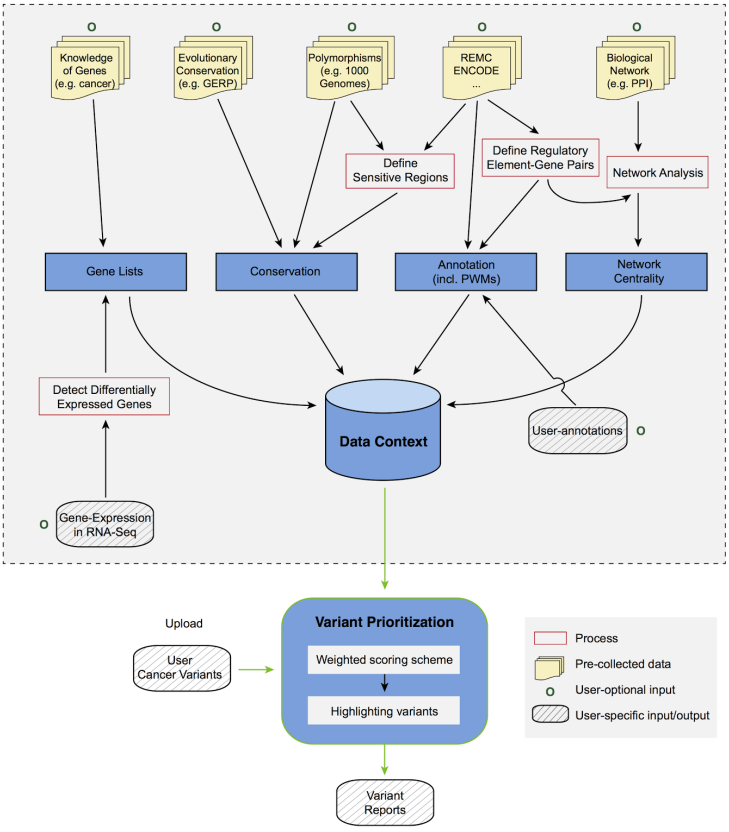


Figure 2. FunSeq2 workflow

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.

Experience in cancer genome analysis consortia

We have participated extensively in consortium analysis of cancer genomes as part of The Cancer Genome Atlas (TCGA) and the Pan Cancer Analysis of Whole Genomes (PCAWG) groups. 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 have developed tools for somatic variant calling \cite{26381235}. We have also extensively used TCGA RNA-Seq data in the development and application of 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 cancer genomes from the PCAWG consortium with ENCODE data to provide a resource for studying noncoding variants in cancer.

B. Research plan

MORE ENCODE

We will prioritize both coding and noncoding mutations for a set of genes of interest. We will first identify putatively important genes in prostate cancer through a combination of recurrence analysis and biological network analysis. We will then functionally prioritize both coding and noncoding mutations for this set of genes.

Definition of a compact annotation for variant analysis

We will first focus on identifying non-coding regulatory regions and linking them to genes by constructing "extended gene neighborhoods". Specifically for enhancers, we will define a compact list through an ensemble method: enhancer candidate identification by integration of pattern recognition based algorithm on ChIP-seq and DNase-seq signals and STARR-seq pipeline, enhancer-target linkage prediction using JEME method, and then filter through high resolution Hi-C experiments. We will also extract the cis-acting TF and RBP binding sites and incorporate them into the extended genes. As with the exon regions within genes, a natural consequence of this is 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.

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). TFs in different levels of the hierarchy reflect the extent to which they directly regulate the expression of other TFs \cite{25880651}. When comparing the common regulators in approximately matched tumor and normal regulatory networks, rewiring (i.e., target changing) analysis may help to identify cancer-associated deregulation. 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 membership and topic models, which used a mixed-membership model to look more abstractly at local gene neighborhoods to re-rank the TFs.

Identification of recurrently mutated elements & genes

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.

We also propose a Negative binomial regression based Integrative Method for mutation Burden analysis (NIMBus), which first intuitively treats mutation rates from different individuals as random variables with a gamma distribution, and resultantly models the pooled mutation counts from a heterogeneous population 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 to create a covariate matrix to predict the local mutation rate with high precision through regression. In addition, it also customizes the most comprehensive noncoding annotations from ENCODE to facilitate interpretation of results. This integrative approach will enables us to effectively pinpoint mutation hotspots associated with disease progression and to better understand the biological mechanisms therein.

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, FunSeq2 score, and an ensemble method that combines scores from many tools that score the functional impact of coding variants \cite{GERP, FunSeq2, 24453961} In addition to the above general scores for coding variants, 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 additional scores within protein repeat regions \cite{27939289}.

Functional prioritization of noncoding mutations

We will first use Funseq2 \cite{25273974} to annotate and score the predicted molecular impact of each variant, including SNVs in the pan-cancer dataset. A closer inspection of the pan-cancer impact score distributions for non-coding variants demonstrate three distinct peaks, which indicates a multimodal distribution of functional impact of non-coding variants (Figure 3). The upper and the lower extremes of this distribution correspond to traditional definitions of high-impact putative driver variants and low impact neutral passengers respectively. In contrast, the middle peak in the intermediate molecular functional impact regime corresponds to what we term impactful nominal passengers. This intermediate functional impact category could include undiscovered drivers (strong & weak) as well as potentially deleterious passengers.

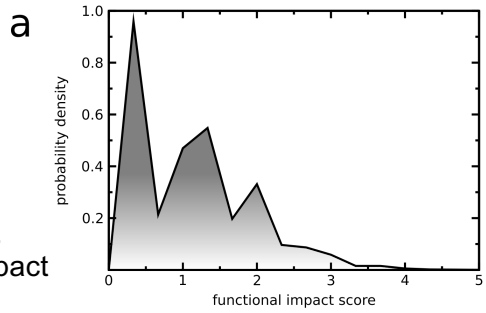


Figure 3. Distribution of FunSeq2 score in Pan Cancer Analysis of Whole Genomes dataset

To integrate the various features, we will expand the weighting system in FunSeq\cite{2409274646} and Funseq2\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, we calculate the probability that it overlaps with common polymorphisms. We 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

*** insert text from paper E showing the multimodal distribution of functional impact scores
We've already observed that functional impact score distribution is complex... Done

Modify variant prioritization for both & noncoding based on allelic activity

Allele-specific variants(ASVs) potentially provide a most direct readout of the functional impact of a variant. We have previously defined allelic elements throughout the genome by conducting a survey of ChIP-Seq and RNA-Seq experiments conducted on 1000 Genomes Project individuals. Gene expression and protein binding are sensitive to mutations in these regions. Our scoring system takes into account not only enrichment of allelic variants within a given 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 up-weighted according to their scores.

Parameter tuning after experimental validation

Let () represent the initial feature parameters chosen at random, where is the number of features. will be optimized using an iterative learning scheme by incorporating new experimental information produced in Aims 2. Because of the high throughput of iSTARR-seq, our strategy is to implement for the first time a iterative learning scheme : the first stage initial learning, the second stage real-time experimental parameter optimization, and the third stage final assessment.

In the first stage, we will randomly select ~500 driver gene as defined by recurrence analysis, PCAWG and TCGA. We will first generate the WT clones of these genes and promoters using xxx-seq. Then, we will select 2 coding variants in coding region and 2 non-coding variants from the promoter region on each gene and generate all ~2,000 variant clones through Clone-seq. Their effects on coding and non-coding variants will be quantified by InPoint and iSTARR-seq pipeline respectively. Starting from the initial tuned , we tune according to the results of ~2000 variants in the first stage. For a specific variant , we define as Bernoulli distributed random variable with indicates that is functional. The expectation of can be predicted through a logistic regression:xxxx (are scaling parameters). To update with experimental validation results , we implement Bayes' rule:. We will use MCMC (Monte Carlo Markov Chain) sampling to search over the parameter space and find the most probable xxx . We will predict the functional impact of all noncoding variants genome-wide, .

In the third stage of final assessment, we will select xxx variants (400 with predicted high impact, 200 with medium impact, and 400 with low impact) on previously cloned driver genes. We will measure their impact on xxxx activities quantitatively through xx-seq.

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

MOVE TO GENERAL SIG SECT.

a. SIGNIFICANCE AND PREMISE

a.3. The importance of investigating functional relevance of coding variants through protein interactome networks

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.

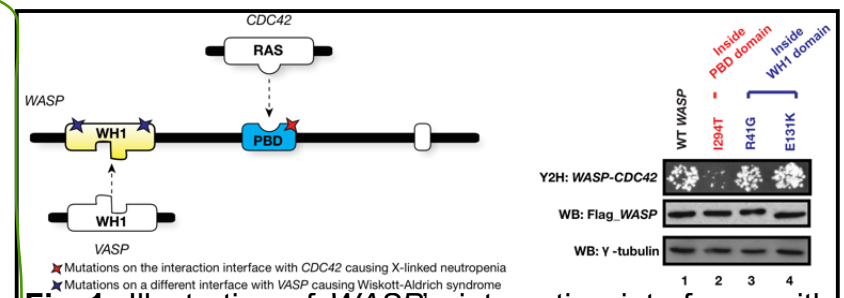


Fig 1. Illustration of *WASP*'s interaction interfaces with *CDC42* and *VASP* and effects on the *WASP-CDC42* interaction by mutations on *WASP*. Mutations causing two distinct diseases are located on two separate interaction interfaces and disrupt different interactions as described in our ***Nature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

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 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.**

b. INNOVATION

b.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

EARLIER.

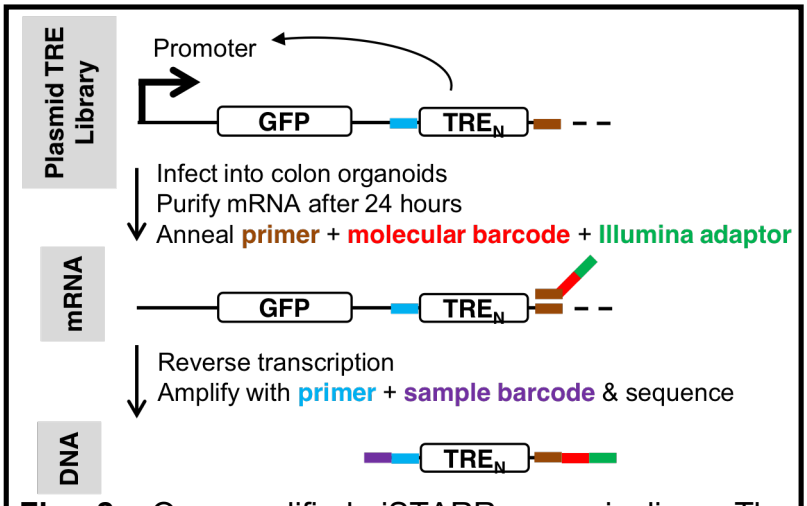


Fig 2. Our modified iSTARR-seq pipeline. The random molecular barcodes (red) uniquely label each mRNA molecule produced before the amplification step during the sequencing library preparation,

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).

b.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).

b.5. Results of our InPOINT assays are physiologically relevant *in vivo*

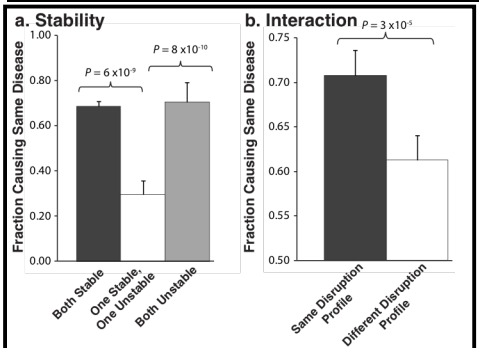


Fig 3. Mutations with similar molecular phenotypes measured by our InPOINT assays tend to cause the same

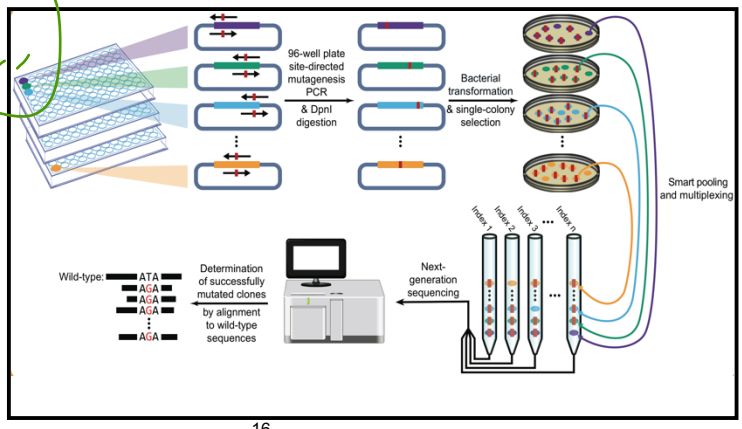
In our previously published ***PLoS Genetics*** study[12], we examined 204 disease-associated mutations using our high-throughput InPOINT pipeline. We find that pairs of mutations in the same genes that are either both stable or both unstable cause the same disease in 68% and 70% of cases, respectively. However, pairs comprising one stable and one unstable mutation cause the same disease in only 30% of cases (**Fig. 3a**) Furthermore, when pairs of mutation disrupt the same set of interactions (i.e., same disruption profile) are significantly more likely to cause the same disease than those that do not (**Fig. 3b**). Overall, these results confirm that the molecular phenotypes measured by our high-throughput InPOINT pipeline are biologically relevant *in vivo*. Furthermore, by comparing the molecular phenotypes, in particular the interaction disruption profiles, of SNVs to those of known disease mutations, potential candidate mutations for a variety of diseases can be identified[12].

c. APPROACH

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.



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

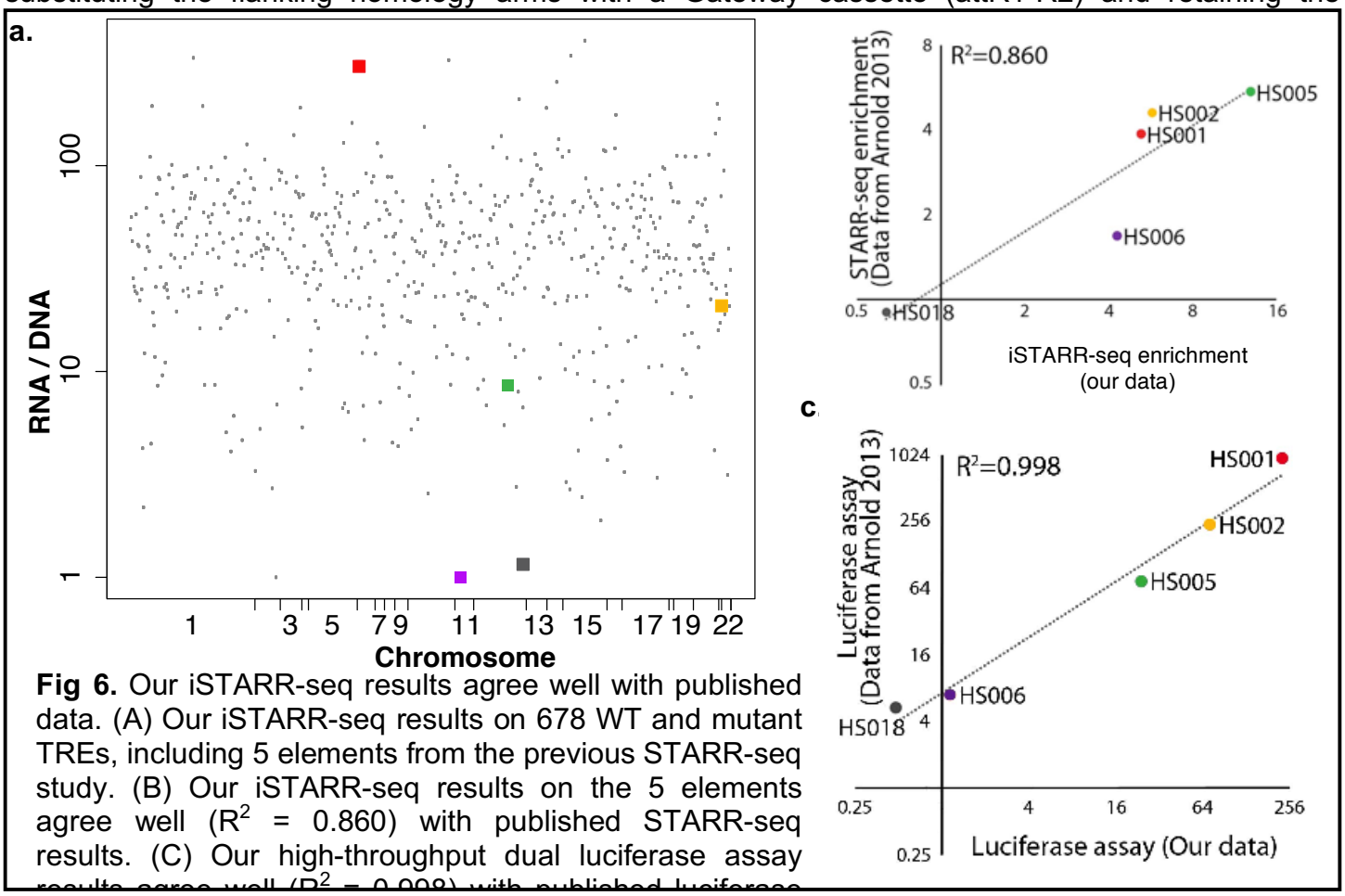


Fig 6. Our iSTARR-seq results agree well with published data. (A) Our iSTARR-seq results on 678 WT and mutant TREs, including 5 elements from the previous STARR-seq study. (B) Our iSTARR-seq results on the 5 elements agree well (R^2 = 0.860) with published STARR-seq results. (C) Our high-throughput dual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

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).

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

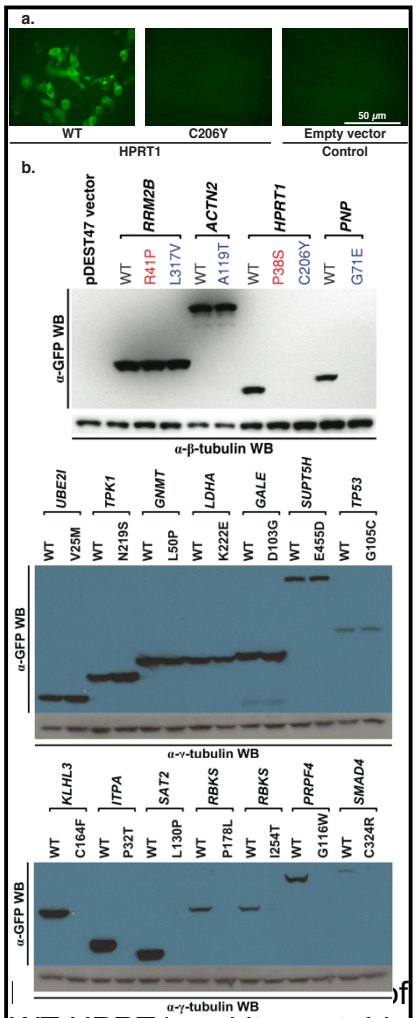
western blotting for 10 random mutants that are stably expressed and 10 random mutants with significantly diminished expression levels (**Fig. 7b**). All western blotting results agree perfectly with our GFP readings[12].

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].

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.



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

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 specific mean I_{rb} specific Z_g and $S_{WT} = \left(\frac{I_g - I_{gb}}{I_r - I_{rb}}\right)_{WT}$ *and* $S_{mut} = \left(\frac{I_g - I_{gb}}{I_r - I_{rb}}\right)_{mut}$ pcDNA-DEST47 vector (with a plate- and s.d. σ_{rb}) are measured. A plate- 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:

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.**

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).

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

A. **Medium-throughput *in vivo* quantification of cellular phenotypes**

[[Still waiting for section from Andre Levchenko]]

B. **Validation of 10 coding and non-coding variants in prostate organoids**

Background for Aim 4

We have recently described our Precision Medicine program and our ability to develop patient derived organoids. In that study, we presented data on 56 tumor-derived organoid cultures, and 19 patient-derived xenograft (PDX) models established from the 769 patients enrolled in an IRB approved clinical trial at Weill Cornell (REF). This expertise has given Dr. Rubin's group over 3 years experience in working with organoids. The focus of this Aim will be on benign prostate organoids that will be derived from patients undergoing surgery in Bern, Switzerland. Since moving to Bern, Dr. Rubin has set up active collaborations with the Urology group, led by George Thallman (see letter of support) and Marianna Di Julio (see letter) to start developing benign prostate organoids for this proposal.

Large-scale drug screens of cell line panels - such as the NCI60 by the National Cancer Institute or the Cancer Cell Line Encyclopedia (CCLE) - have addressed compound sensitivity in cancer cells to identify mechanisms of growth inhibition and tumor-cell death [1-3]. A more recent study of pharmacogenomic interactions in cancer links genotypes with cellular phenotypes with the purpose of targeting select cancer sub-populations [4]. Unfortunately, for many cancer types, traditional cell culture methodologies do not adequately model the biology of the native tumor. The high failure rate of preclinical compounds in clinical trials clearly demonstrates the limitations of existing preclinical models [5,6]. The accuracy of *in vitro* drug screens is therefore dependent on the optimization of cell culture tools that more closely mirror patient disease.

For this Aim, we propose using Organoid technology as an intermediate model between in vitro cancer cell lines and xenografts as shown for colorectal, pancreatic and prostate cancer [7-12]. This technique 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 [13-20].

Preliminary Results

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.

Approach

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.

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 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)**

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).