D-3 Approach Aim 3 - Medium scale validation of the prioritized variants

D-3-a Preliminary results related to validation

D-3-a-iv Performance, throughput, and cost of our Clone-seq pipeline

To set up our Clone-seq pipeline, we first focused on 27 interactions that can be detected by our version of Y2H, are represented in co-crystal structures, and have known missense disease mutations on their corresponding proteins in HGMD. Of these 27 chosen interactions, 24 have disease mutations on the corresponding interaction interfaces and 15 have mutations away from the interfaces. For interactions that have more than one mutation on and/or away from the interfaces, we randomly picked one for each interaction. To generate these 39 mutant alleles, we picked 4 colonies for each mutation. As a reference, we also pooled together all the WT alleles in our human ORFeome library to be sequenced together with the 4 pools of the mutagenesis colonies. In total, there are 40.1 million Illumina HiSeq 1×100 bp reads for our Clone-seq sample.

These reads were then de-multiplexed and mapped to the genes of interest using the BWA aln algorithm. There is an average of > 2,500× coverage at all desired mutation sites. For each allele of interest, we identified all reads that map to the position of the mutation of interest (Rall) and those that actually contain the desired mutation (Rmut). We then calculated a normalized score that quantifies the fraction of reads that contain the desired mutation:

$$S = Rmut/Rall \times 1/k$$

where k is the number of different mutations for the same gene.

Out of 156 colonies containing the 39 mutations, 125 of them were successful. Thus, our overall PCR-mutagenesis success rate is 80%. In fact, we were able to pick correct clones for all 39 mutant alleles using only the first two pools in Clone-seq. All 78 clones from the first two pools, from which the correct ones used in subsequent steps were selected, were Sanger sequenced for verification. All 55 Clone-seq positive results with S > 0.8 were confirmed, and there is a clear separation in the S scores between the successful and failed clones (Fig. 4). One major advantage of our Clone-seq pipeline is that we can now carefully examine whether there are other unwanted mutations introduced during the PCR process. We found that there are on average 4-5 additional mutations introduced in each pool of the 39 colonies. This corresponds to a 0.013% error rate, in agreement with previous studies. The detection of additional mutations, especially those far away from the mutation of interest, cannot be achieved with the traditional site-directed mutagenesis pipeline using Sanger sequencing. These unintended mutations could lead to erroneous downstream results.

Table 1. Cost comparison between Sanger and Illumina sequencing1.

Traditional Sanger sequencing		Clone-seq	
Unique mutations	3,047	NEBNext Multiplex Oligos (E7335S)	$19.80
Colonies per mutation	4		
Total number of samples	3,047x4=12,188	NEBNext DNA Library Prep Master (E6040S)	$105

Re-sequencing needed2	5%		
Number of 96-well plates needed	137	Illumina HiSeq, single-end, 100 bp sequencing lane	$1,175
Cost per plate	$300		
Minimum cost3	43x300=$12,900	Total cost	$1,299.80
Total cost	137x$300=$41,100		

1All costs are based on internal Cornell pricing.

2Sanger sequencing has an average failure rate of 5%.

3The minimum cost is the least amount of money spent in Sanger sequencing the expected number of samples needed to obtain one correct clone for each mutation of interest.

For our Clone-seq samples, we obtained only 40.1 million reads out of a total of 125 million reads in a single lane of a 1×100 bp HiSeq run with >2,500× coverage. However, to determine S to a least count of 1%, we only need 100× coverage. Since the separation between a successful mutagenesis attempt with the lowest S and an unsuccessful mutagenesis attempt with the highest S is 0.28, 100× coverage makes this separation >25 times our least count. We further increase this separation to >60 times our least count by requiring S > 0.8 for a mutagenesis attempt to be considered successful. 100× coverage is also sufficient for a conservative variant calling pipeline to identify additional unwanted mutations with high confidence35,36. Thus, we can obtain 39×(125/40)×(2,500/100) = 3,047 alleles with a single lane of a 1×100 bp HiSeq run using the Clone-seq pipeline. Overall, our Clone-seq approach will drastically improve the throughput of site-directed mutagenesis and decrease the total cost by at least 10-fold (Table 1).

To further test Clone-seq, we identified a set of 446 SNVs from the published ESP6500 dataset[36] that are at the interface of protein interactions and are amenable to testing using our high-throughput Y2H approach. Using our Clone-seq pipeline, we performed large-scale, site-directed mutagenesis to generate clones for these 446 SNVs. We sequenced 4 colonies each for the 446 alleles of interest using one full 1×100 bp MiSeq run. We obtained ~14 million reads and aligned them to the reference sequence database using BWA[79]. For each allele of interest, we identified all reads that map to the position of the mutation of interest (R_{all}) and those that contain the desired mutation (R_{mut}). The read coverage surrounding the mutation of interest was ~300× per allele. Using a threshold of $S > 0.8$, approximately 75% of the colonies contain the desired mutation. We were able to choose at least one colony that contains only the desired mutation (without additional unwanted ones) for 437 of the 446 mutagenesis attempts, a success rate of 98.0%.

Overall, our pipeline has been significantly optimized to make it very efficient. We established a web-tool (http://www.yulab.org/Supp/MutPrimer) to design mutagenesis primers both individually and in batch. MutPrimer can design ~1,000 primers for ~500 mutations in one batch in less than one second. All primers for the 476 mutations in this study were generated by MutPrimer. All mutagenesis PCRs are performed in batch using automatic 96-well procedures. Since single colony picking after bacterial transformation of mutagenesis PCR product is a rate-limiting step, we rigorously optimized this step and found that adding 10 μL mutagenesis PCR products to 100 μL competent cells and plating 50 μL transformed cells give the best transformation yield and well-separated single colonies. Furthermore, rather than individually streaking transformed cells onto agar plates one sample at a time, we were able to significantly increase throughput by spreading colonies using glass beads onto four sector agar plates which are partitioned into four non-contacting quadrants. In this manner, a 96-well plate of transformed bacteria can be plated out onto 24 four-sector agar plates in ~15 minutes. Traditional site-directed mutagenesis pipelines require miniprepping each of the selected colonies and sequencing them separately by Sanger sequencing. To drastically improve the throughput of our Clone-seq pipeline, we pooled together the bacteria stock of a single colony for each mutagenesis attempt to perform one single maxiprep, which makes the library construction step much more efficient and amenable to high-throughput. Furthermore, existing variant calling pipelines cannot be applied to our Clone-seq results because the expected allelic ratios built into these pipelines are a function of the ploidy of the organism. However, in our Clone-seq pipeline there is no concept of ploidy. We pool together many mutations for one gene in the same pool (e.g., 40 mutations for *MLH1*) and different genes often have different numbers of mutations. Our *S* score calculation and unwanted mutation detection pipeline was designed according to our pooling strategy .

In total, we have used the Clone-seq pipeline to successfully generate 476 (39 + 437) clones with the desired mutant alleles. The results confirm the scalability, accuracy, and throughput of our Clone-seq pipeline. Through careful considerations, we are confident that this approach can be scaled up to generate the ~1000 SNVs as proposed.

D-3-a-v Reporter luciferase assays confirm validity of in silico TF binding sites

Using an *in silico* approach we determined genome wide distribution of ER in prostate cancer. Intriguingly, we observed a robust recruitment to non-coding genome and identified several intergenic sites that correlated with high ER occupancy. Analysis of recruitment vs transcript profiles confirmed that ER recruitment was associated with productive transcription of long noncoding RNA. Recruitment of ER upstream of NEAT1 lncRNA was addressed in greater details. Reporter assays using promoter luciferase constructs encompassing upstream regulatory regions of NEAT1 and corresponding to two ER binding sites are described in Fig. 9. Interestingly, we discovered that NEAT1 is associated with chromatin and regulates transcription of key prostate cancer genes. Recruitment of NEAT1 was evaluated by ChIP assay and influence on key target genes like PSMA was validated using ChIP and reporter assays (Fig. 10). Functional validation of NEAT1 functions revealed a predominant tumorigenic role as overexpression of NEAT1 was sufficient to augment proliferation, invasion and migratory behavior of prostate cancer cells (Fig. 11).

D-3-b Research plan related to validation

D-3-b-i Overview of validation strategy

Identification of rare variants and understanding the influence thereof on repertoire of biological responses will afford us a unique opportunity to understand causal role of these variations on other somatic mutations associated with diseased states including but not limited to cancer.

We will use Clone-seq to generate ~300 candidate non-coding variant clones identified in Aim 1 and 2. The clones will then be subjected to the downstream reporter assays. Because of the throughput of our Clone-seq and luciferase reporter assays, we will perform iterative learning. That is, we will first clone and test ~150 candidate ncSNVs predicted by our computational learning algorithm. Based on the reporter assay results, we will fine tune the parameters of the learning algorithm, and then perform the predictions again. We will then clone and test another ~150 ncSNVs to confirm the performance of our algorithm. Top candidate ncSNVs that are shown to significantly alter gene expression will be selected for further *in vivo* validations as described in **Aim 4**.

D-3-b-i-(1) ***High-throughput site-directed mutagenesis PCR and E. coli transformation.*** Primers for site-directed mutagenesis are selected based on an optimized version of the protocol accompanying the QuikChange Stratagene site-directed mutagenesis kit (200518). 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. All WT clones are obtained from the Human ORFeome 8.1[81]. PCR products are digested by *DpnI* (NEB R0176L) overnight at 37 °C (30.5 µL PCR product, 3.5 µL 10× NEBuffer 4, 1 µL *DpnI*). *E. coli* competent cells are prepared in 96-well plates with 20 µL cells per well. 10 µL of *DpnI*-digested PCR products are added to the competent cells using the Tecan robot. After heat shock, 800 µL of SOC recovery medium is added to each well using the Tecan robot and the plate is incubated at 37 °C for 1 hr with vibration. A 20 µL aliquot of the cells is then spotted onto LB + Spectinomycin plates in a fully automated fashion using the Tecan robot. The cells are then spread out in the plates through vigorous shaking with glass beads, as is routinely done in the lab. The plates are incubated overnight at 37 °C. The next day, four colonies per allele are picked for Illumina sequencing. We have already carefully titrated the amount of cells plated so that almost all plates have well-separated single colonies.

D-3-b-i-(2) ***illumina library preparation and HiSeq sequencing.*** DNA plasmids from all four colonies of all alleles are mini prepped using our fully-automated 96-well miniprep pipeline. Four libraries representing one colony of each allele are generated according to Illumina protocols and labeled with distinct barcodes. These four libraries are then mixed into one pool for one 1×100 bp HiSeq run. The *S* score for each colony of each allele is calculated as described above. As shown in **Fig. 4**, we found that all clones with $S > 0.44$ are confirmed to be correct via Sanger sequencing with a clear separation between those that are correct and those that are not. However, to ensure that the clones we pick are correct, we require $S > 0.8$ for a colony to be scored as containing the desired mutation.

D-3-b-i-(3) Functional consequences: Reporter assays

Reporter assays that employ either LUC or next generation reporter vectors can provide direct insight to functional relevance of SNPs on target gene. GeneCopoeia offers Gaussia-luciferase (GLuc), eGFP,or mCherry based lentiviral or non-viral promoter reporter clones that can serve as efficient tools to study transcription regulation. Minimal essential promoter region for each WT target gene will be subcloned from germline DNA using TOPO cloning kit (Invitrogen). If patient sample that harbors the mutation is available, we will amplify the corresponding mutant promoter sequence from the genomic DNA of the patient. PCR products will be cloned upstream to pGL-4-LUC promoter reporter plasmid or upstream to Gluc vectors. For each WT DNA Target gene-promoter plasmid a corresponding MT DNA Target gene-promoter plasmid will be generated using site directed mutagenesis utilizing QuikChange Lightning (Agilent). In this way we will have 300 WT promoter plasmids and 300 MT promoter plasmids in both PGL-3 LUC and Gluc background. We will utilize a panel of adherent cell lines. We will use prostate cancer as a model for the validation but we expect that the results will be generalizable to a number of cancers.

Cells will be seeded in 6 well plates and transfected with promoter reporter WT and mutant plasmid constructs. 48 hrs after transfection promoter activity will be measured following manufacturer's instructions. Assay values will be normalized using internal renilla luciferase as control.

Our expectation is that *in vitro* promoter LUC assays will inform us if a particular mutation had any effect on transcription.

D-4-a Preliminary results related to validation:

D-4-a Preliminary results related to validation

D-4-a-i Low-frequency functionally active intronic & intergenic inherited variants predisposing to cancer

Emerging insights into the genetics of constitutional disease etiology demonstrate that germline polymorphisms are associated with a variety of diseases including Alzheimer's, Parkinson's, mental retardation, autism, schizophrenia \cite{19715442}and cancer \cite{19536264,18685109}. Relevant to this proposal our group recently performed a large scale profiling study for 2,000 individuals from the Tyrol Early Prostate Cancer Detection Program \cite{18321314,16829552}cohort. This cohort is part of a population-based prostate cancer-screening program started in 1993 and intended to evaluate the utility of intensive PSA screening in reducing prostate cancer specific death. By genotyping DNA extracted from peripheral blood samples, we annotated the cohort on more than 5,000 CNVs and 900,000 SNPs and then queried inherited low frequency deletions variants \cite{20059347} for their impact in driving prostate cancer \cite{20479773} and the more aggressive form of the disease \cite{10351184}. We reported on coding and non-coding functionally active risk variants. Among the top hits of the case-control study, an intronic variant in the *Alpha-1,3-mannosyl-glycoprotein 4-beta-N-acetylglucosaminyltransferase C (MGAT4C)* demonstrated transcript abundance association with genotype states both in prostate and in lymphoblastoid cells, significant increase in cell and migration upon overexpression in benign and cancer prostate cell lines, and significant decrease in proliferation upon knock down of *MGAT4C* expression with siRNA. In addition, we suggested that intergenic PCA risk variants affect gene regulation through modified transcription factor binding activity of the Activator Protein 1 (AP-1) \cite{20299548,21862627}. Altogether, we demonstrated that inherited variants may directly or indirectly modulate the transcriptome machinery of known oncogenic pathways in prostate cancer facilitating carcinogenesis.

D-4-a-ii In vitro characterization of SNPs within enhancer elements bound by AR and/or ER

The Tyrol Early Prostate Cancer Detection Program cohort is a well characterized cohort with centralized data collection that ensures proper patients' follow-up annotations and availability of well-preserved tissues and blood samples. The cohort currently includes more than 3,000 men. As part of our Trento-Innsbruck-Cornell collaboration, we further studied the genetics of prostate cancer individuals coupling serum levels and genomics data. Specifically, we studied the impact of genetic variants relevant to the metabolism of Dihydrotestosterone \cite{20056642}(DHT), the most potent form of androgen, and investigated the incidence of common genomic rearrangements with respect to PSA levels and age at diagnosis \cite{23381693}.

It has been shown that a significant fraction (26%-35%) of inter-individual differences in transcription factor binding regions coincides with genetic variation loci and that about 5% of transcripts levels are associated with inherited variant states \cite{20299548}. Genotype-transcript associations have been reported at large for multiple types of inherited variants \cite{21479260,20220756,20220758,21862627,17289997}, however experimental evidence of inherited variants allele-specific effect on enhancer activity are lacking. In order to study the potential role of inherited genetic variants within regulatory elements in the context of hormone dependent human, we have performed an unbiased computational search for AR/ER bound enhancers elements containing SNPs followed by *in vitro* characterization of selected variants. **Table 1** shows counts of SNPs from the dbsnp137 set within AR \cite{20478527} and/or ER (Chakravarty D, *submitted*) binding sites that intersect peak ENCODE data \cite{22955616} generated from 20 cell-lines and ChIP-seq experiments for H3K4m1, H3K4me1+H3K4me3, H3K9ac, H3K27ac, Dnase-seq and FAIRE-seq. For each marker the consensus was generated as the merge of all the regions that are present in at least 2 cell lines and comply with a set of filters. **Fig. 8** shows examples of AR-responsiveness and SNPs impact on putative enhancer elements in MCF7 cells (Garritano S, Demichelis F, *unpublished*).

D-4-a-iii Modeling mutations in cell lines using CRISPR CAS system:
Mutation in the MAP3K7 gene is seen in castrate resistant prostate cancer patients. Inorder to determine the functionality of the mutation we have used the CRISPR CAS system to generate the mutation in cell lines. We have successfully introduced cancer-specific MAP3K7 mutation in VCaP cells using the CRISPR-CAS system. Sequencing of cell lines confirmed mutation. Next we studied the genomic influence of MAP3K7 mutation in evolution of castrate resistant prostate cancer. Another example is the deletion of the FANCA gene evidenced in metastatic prostate cancer patients. We have used the CRISPR CAS system to generate FANCA deletion in prostate cancer cell lines..

Briefly, the CRISPR/Cas9 plasmid (Px459) was obtained from Addgene (Cambridge, MA). Using Ran *et al*(15) protocol we identified a FANCA CRISPR DNA target sequence using algorithms based on analysis in Hsu *et al*(16). The corresponding oligonucleotides were ordered (IDT Coralville, IA) and were cloned into Px459 vector. Sanger sequencing confirmed integration of the FANCA target site into the vector.

DETAIL

D-4-a-iv Validation and functional evaluationof physiologic role of somatic mutation predicted by FUN-seq bioinformatics pipeline. Mutation in RET promoter was determined using insilico FUN-seq pipeline. Bioinformatic analysis using FUN-seq pipeline predicted gain of AP1 motif in promoter of RET promoter. Using a luciferase based reporter assay we studied the promoter activity of WT and mutant RET promoter in LnCaP and DU145 cell lines. Luciferase activity confirmed that mutant promoter was X fold active than the WT promoter. Further addition of

AP1 inhibitor compromised the activity, indicating that the observed increase in promoter activity was indeed due to AP1 binding at promoter elements.

Approach:

D-4-b-i Targeted genotyping: We will determine if any or all 10 variants selected based on successful validation in Aim 3 are associated with cancer or cancer causing characteristics. We will achieve this by studying the specific variant in test cohort. We will use both the Tyrol cohort (described above) and the Early Detection Research Network (EDRN) \cite{0000005} prostate cancer cohort with thousands of prostate cancer individuals as well as normal controls. The prostate cancer cohort include men enrolled at three sites as part of the Prostate Cancer Clinical Validation Center that prospectively enroll individuals at risk for prostate cancer at Beth Israel Deaconess Medical Center (Harvard), at the University of Michigan (Michigan) and at Weill Cornell Medical College (Cornell). Cases are defined as men diagnosed with prostate cancer and controls are men who have undergone prostate needle biopsy without any detectable prostate cancer and no prior history of prostate cancer.
We will first take the highest prioritized variants then subject them to validation. Overall we plan to start the validation pipeline with the top ~10 elements identified from the reporter assays (as described above). TaqMan assays for these 10 variants will be performed on 4,000 cases to see if the precise variants recur in a larger cohort. From this group, we will select top third of the variants (~6), based on recurrence, that we will follow up for detailed functional screening, to be discussed below. This functional screening will be through various reporter assays (e.g. luciferase) looking for the effect on the target gene and also from using the CRISPR/Cas system. For controls, we will utilize deeply sequenced control cohorts (individuals with no cancer) that are already available, including deeply sequenced trios from the 1000 Genomes Project \cite{0000006}, 500 individuals with Complete Genomics sequencing also from 1000 Genomes \cite{0000007} and non-cancerous individual from the UK10K project \cite{0000008}.

Superior allelic discrimination is achieved in these assays as they utilize TaqMan minor groove-binding (MGB) probes. This technique generates a low signal to noise ratio and affords a greater flexibility. The Taqman probes are functionally tested to first ensure assay amplification and optimization for amplification conditions.
Methods: Genomic DNA will be extracted from the blood cellular-EDTA samples in a high-throughput fashion using the QIAamp 96 DNA Blood Kit (Qiagen). All DNAs are evaluated by NanoDrop spectrophotometer (NanoDrop, Thermo Scientific) and gel electrophoresis (2% agarose). For TaqMan Real-Time Quantitative PCR, each DNA sample will be diluted to 10 ng/ml with nuclease-free water.

D-4-b-ii Evaluation of functional consequence of variants

Based on the Taqman results, we will pick the top ~6 variants for functional follow up.

D-4-b-ii-(1) Functional consequences: RNA-seq

We have RNA sequencing data for 85% of the individuals enrolled in the cohort. To fill out the dataset, RNA sequencing will be completed on the remaining where we see recurrent variants (on up to ~160 individuals). The RNA-seq will be done according to the protocols in \cite{21036922}. This analysis will inform us if a SNP (in promoter or enhancer regions) has any effect on transcription of target gene. This analysis will provide a comprehensive list of SNPs that might correlate with loss or gain of expression. Recurrent rare SNPs will be further validated by PCR assays using primers that can amplify the genomic region encompassing the SNP. PCR will be followed by direct sequencing of amplicon using an ABI 3730 DNA Sequence Analyzer on a subset of tumor-normal pairs to verify the individual promoter/enhancer mutations for further confirmation.

D-4-b-ii-(2) Functional consequences: CRISPR/CAS system

We will utilize the newly discovered CRISPR/CAS system \cite{0000009} to generate endogenous mutations in target genes in a panel of prostate cancer cell lines (VCaP, LnCaP, DU145 and PC3). This unique system will provide us an opportunity to directly modulate endogenous genes and minimize artifacts due to the transfection based reporter assays. Using CRISPR/CAS mediated genome-engineering method \cite{23643243} we will directly generate mutations within promoter/enhancers of target genes. Theoretically we generate 6 individual SNPs in each cell line and will study functional relevance of these changes compared to WT. In case of rare mutations, which occur within both promoter and enhancer regions of the same gene, we will develop cell lines having these combinatorial mutations.

Mutations within regulatory regions like promoter and enhancer regions might contribute to one or more biological effects as described in the schematic (Fig. 12). In addition to loss or gain of cognate coding transcript, it is quite conceivable that the SNPs might alter expression of non-coding transcript. To capture the complete influence of rare nominated SNPs at genomic and transcriptomic level we will perform RNA seq. The schematic (Fig. 12) shows representative iterations of plausible genomic changes that will be captured in this validation.

For modeling mutations in non coding RNA, prostate cell lines will be screened for the expression of the non coding RNA, and in cells having a high endogenous expression of the ncRNA, CRISPR/CAS system will be used to generate the mutation.

D-4-b-ii-(3) Functional consequences:

The mutant and WT cell lines generated using CRISPR/CAS system will be monitored for a) phenotypic changes by confocal microscopy and actin staining to determine effects of mutation on cytoskeletal reorganization b) Influence on proliferation by MTT and CellTiter-Glo® Luminescent Cell Viability Assay (Promega) c) Influence on invasive and migratory potential using, matrigel coated invasion and boyden chambers in 24 well format d) senescence by Bgal staining e) apoptosis by tunnel assay.

D-4-b-ii-(4) Functional validation of mutation in non coding RNA:

Total RNA will be extracted from cell lines expressing the wild type and the mutant ncRNA and RNA sequencing will be performed to determine the mutation specific gene signature.

D-4-b-ii-(5) Effect of the mutation on TF binding

In vitro EMSAs will confirm specific binding to WT or mutant sequence by a particular transcription factor.

EMSA (electrophoretic mobility shift assay) is a common technique employed to study protein-DNA interaction. We will use the WT and the MT sequences to determine binding to a transcription factor predicted to be present at the site of mutation.

Chromatin immuno-precipitation (ChIP) assays for TFs overlapping the variant will be conducted to determine if the variant can distort TF binding. This would help validate the variants that are predicted to be motif breakers. Alternatively for the SNVs predicted to create a new motif, ChIP experiments will help validate binding.