UNIVERSITY OF CALIFORNIA, LOS ANGELES **UCLA**

BERKELEY • DAVIS • IRVINE • LOS ANGELES • MERCED • RIVERSIDE • SAN DIEGO • SAN FRANCISCO

SANTA BARBARA • SANTA CRUZ

Sriram Kosuri, Sc.D.
Assistant Professor
Dept. of Chemistry and Biochemistry
PO Box 951569
607 Charles Young Drive, East
Los Angeles, CA 90095-1569

October 18, 2013

To Whom It May Concern:

Please find attached my application for the 2014 NIH New Innovator Award (RFA-RM-13-007) entitled "Reverse Genomics of Regulatory Elements Governing Splicing". I am a new Assistant Professor in the Department of Chemistry and Biochemistry at UCLA and my appointment begins January 1st, 2014. I am routing this under the following Agency codes: Primary, 8: High-Throughput and Integrative Biology; Secondary, 9: Quantitative and Computational Biology.

This proposal describes existing collaborations for obtaining synthetic oligo and gene libraries. These collaborations include Agilent Technologies (Contact: Steve Laderman, Director, Molecular Tools Laboratory, Agilent Technologies), CustomArray (Contact: Marcelo Caraballo, Senior Scientist, CustomArray, Inc.), Gen9 (Contact: Kevin Munnelly, President and CEO, Gen9), and the Wyss Institute (Richard Terry, Lead Senior Staff Scientist, Wyss Institute). Please feel free to contacting these individuals or me directly if there are any questions.

Thank you,

Sriram Kosuri
skosuri@chem.ucla.edu
cell: 617-852-4806

Any individual's human genome will have several million deviations from the consensus human genome sequence. Understanding whether these mutations are relevant or important is difficult because most mutations are uncommon and we do not have perfect understanding of how sequence affects function. Here we will develop new high-throughput methodologies to understand how and the extent to which this genetic variation affects function to develop better genetic diagnostics and therapeutic interventions.

Reverse Genomics of Regulatory Elements Governing Splicing

Differences in our individual genomes give rise to most of human diversity. A decade removed from the Human Genome Project, much of how the genome directs phenotype still remains a mystery. Consortiums like ENCODE seek to identify functional DNA elements in humans and other model organisms by correlating functional outputs with sequence using genome-wide data sets. However, these studies do not necessarily improve our ability to interpret how DNA elements act in new contexts or when mutated. Such an understanding will be critical to predict the effects of sequence alterations on phenotype and to engineer biology for future medicinal or technological purposes.

Combinations of DNA elements act as codes controlling particular functions like transcription, splicing, localization, and silencing. Deciphering these codes is difficult, as the limited set of natural variants is typically insufficient to control for variables such as sequence composition or element combinations. Proving that particular sequences have causative effects on gene expression requires carefully controlled reverse genetic studies. Conducting such experiments on genome-wide scales is difficult because of our inability to (1) rapidly alter the sequence and context of individual genetic elements and (2) quantify the consequences of thousands of such changes.

My central vision is to decipher cis-regulatory codes controlling gene expression by scaling reverse genetics experiments to genomic scales using multiplexed measurements of defined synthetic DNA libraries. I will build upon my work developing next-generation gene synthesis technologies and multiplexed reporter assays to systematically determine how sequences governing mammalian gene expression act in concert by doing thousands of controlled experimental tests simultaneously.

Here, we will apply these technological developments to study how genetic regulatory elements control the process of pre-mRNA splicing. The major sequence elements controlling splicing, namely the splice donor, acceptor, and branch sites, do not convey enough information to specify exon inclusion or exclusion alone. Other regulatory elements, such as exonic or intronic splicing enhancers and suppressors, are known to affect splicing in a complex code that can vary based on tissue or cell type. We will systematically interrogate and refine the splicing code by leveraging the new technological developments proposed here. Studying splicing will help focus development of a complete suite of tools and technologies, which will later let us attack other forms of cis-elements controlling gene regulation.

Project Science Areas: 8 HIB; 9 QCB

REVERSE GENOMICS OF REGULATORY ELEMENTS GOVERNING SPLICING

A. PROJECT DESCRIPTION

Differences in our individual genomes give rise to most of human diversity. A decade removed from the Human Genome Project[1,2], much of how the genome directs phenotype still remains a mystery[3,4]. Consortiums like ENCODE seek to identify functional DNA elements in humans and other model organisms by correlating functional outputs with sequence using genome-wide data sets[5-7]. However, these studies do not necessarily improve our ability to interpret how DNA elements act in new contexts or when mutated. Such an understanding will be critical to predict the effects of sequence alterations on phenotype and to engineer biology for future medicinal or technological purposes.

Combinations of DNA elements act as codes controlling particular functions like transcription, splicing, localization, and silencing. Deciphering these codes is difficult, as the limited set of natural variants is typically insufficient to control for variables such as sequence composition or element combinations. Proving that particular sequences have causative effects on gene expression requires carefully controlled reverse genetic studies. Conducting such experiments on genome-wide scales is difficult because of our inability to (1) rapidly alter the sequence and context of individual genetic elements and (2) quantify the consequences of thousands of such changes.

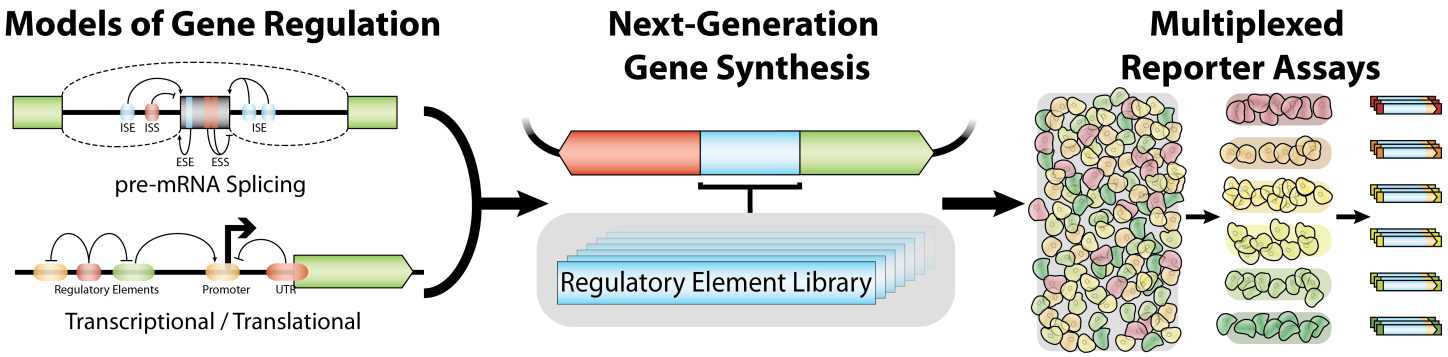

Figure 1. Reverse Genomics. We will bring reverse genetics experiments to the genomics age by developing methods to simultaneously perform thousands of reporter experiments to improve models of gene regulation.

My central vision is to decipher cis-regulatory codes controlling gene expression by scaling reverse genetics experiments to genomic scales using multiplexed measurements of defined synthetic DNA libraries (Fig. 1). I will build upon my work developing next-generation gene synthesis technologies and multiplexed reporter assays to systematically determine how sequences governing mammalian gene expression act in concert by doing thousands of controlled experimental tests simultaneously.

Here, we will apply these technological developments to study how genetic regulatory elements control the process of pre-mRNA splicing. The major sequence elements controlling splicing, namely the splice donor, acceptor, and branch sites, do not convey enough information to specify exon inclusion or exclusion alone. Other regulatory elements, such as exonic or intronic splicing enhancers and suppressors, are known to affect splicing in a complex code that can vary based on tissue or cell type[8,9]. We will systematically interrogate and refine the splicing code by leveraging the new technological developments proposed here. Studying splicing will help focus development of a complete suite of tools and technologies, which will later let us attack other forms of cis-elements controlling gene regulation.

This proposal outlines methodologies that will require new developments in chemistry (emulsions), molecular biology (gene synthesis), genetic engineering (genome editing), genomics (next-generation sequencing), and bioinformatics (design and analysis). As a new investigator beginning my lab in January 2014, the New Innovator Award will give me the opportunity to focus on technological developments that will not only have broad impact, but also allow me to transition to traditional hypothesis-driven funding mechanisms as these technologies mature. My training, while unconventional, makes me uniquely qualified to develop these technologies and apply them to new disciplines.

The remainder of this proposal is structured as follows: First, I discuss how cis-regulatory elements (CREs) govern pre-mRNA splicing in humans and why predicting splicing patterns based on sequence is still difficult (**Background)**. Second, I summarize my previous work on developing reverse genomic approaches to quantify how CREs affects bacterial expression, which helps set the stage for this proposal (**Previous Work**). Third, I outline the technological developments required in gene synthesis, genome engineering, multiplexed assays, and informatics to move from bacterial to mammalian systems (**Technology Development**). Fourth, I discuss how we will apply these technological developments to better understand how CREs direct pre-mRNA splicing in human cell lines (**Pre-mRNA Splicing**). Finally, I will discuss the proposal's potential impacts and innovations, and my own background (**Impact, Innovativeness, Candidate Background**).

Background

The average human gene will undergo 10 RNA splicing events that remove introns from pre-mRNA to bring together the exons that code for the protein sequence. The spliceosome is a loosely associated complex of over 100 proteins and 5 small nuclear ribonucleoproteins (snRNPs) that directs splicing[10]. Splicing can act as a point of regulation as the level of splicing for a particular exon can alter a protein's quantity and activity. Indeed, splicing is often tissue specific in that particular exons can be alternatively included or excluded depending on the cell type in which the transcript is expressed[11]. It is estimated that up to 95% of multi-exon human genes are alternatively spliced[12,13]. Numerous studies have shown that even point mutations can cause splicing dis-regulation that can be causative for a variety of human diseases[14-16].

Various lines of genetic studies indicate that the spliceosome first recognizes exons and their boundaries in a process called exon definition[17,18]. For example, mutations to splice sites in an exon most often lead to skipping of the entire exon rather than retaining an intron[19,20]. The major sequence determinants of splice sites are directly located at the splice junctions; the splice acceptor and donor have 9 and 15 base-pair (bp) consensus sequences located on the 5' and 3' end of the exons respectively[21-23]. In addition, the branch point sequence that plays a role during splicing has a highly degenerate 7bp sequence ~33bp upstream of splice acceptor[24-26]. The combined degeneracy of all these sequences results in many more possible splice sites and false exons in a given pre-mRNA than are utilized in vivo[27,28]. Numerous computational and *in vitro* studies have shown that other cis-regulatory elements are required to distinguish false splice sites from real ones[29]. These sequences are short motifs that are broadly classified as exonic splicing enhancers (ESE) and suppressors (ESS) as well as intronic counterparts (ISEs & ISSs). How the complex interaction of these sequences and other factors combine to direct splicing has been collectively termed the "splicing code"[8,30-34].

Numerous computational analyses have identified many regulatory motifs that are associated with increased or decreased splicing efficiencies[35-38]. In fact, so many motifs have now been identified in the literature (>75% of all sequence space) that they often do not significantly improve our capacity to differentiate pseudo-exons from true exons[32]. More recent computational studies have relied upon finding interacting motifs that cooperate to affect splicing efficiencies to generate more predictive splicing models[8,39,40]. However, predicting the quantitative effects of regulatory elements on splicing still remains an elusive goal[34]. One potential reason is that computational models rely on observational data because large reverse genetic studies on the effect of regulatory elements on splicing efficiency are lacking. In addition, alternative splicing is known to play a significant role in differential gene expression in different tissue types[8,41-43]. Splicing differences between cell types are thought to occur due to differential regulation of splicing machinery[33]. However for the most part, the CREs that lead to these functional changes have only been examined by correlation to tissue-specific RNA-seq that are an average of many different cell types.

Understanding this splicing code is important to understand how to interpret variation and mutation in our genomes. For example, each newly sequenced human genome contains several million deviations from the reference genome sequence, most of which have unknown function[44-46]. Within genes, SNPs are twice as likely to occur within introns than in exons[44]. Even then, synonymous mutations within exons can cause splicing defects and pathological disorders[47,48]. More generally, the splicing code is one of many such cis-regulatory codes that control a variety of processes such as transcription, translation, mRNA degradation and epigenetic modifications. Thus, as we evaluate new genome sequence, how are we to understand the effects of such variation? Can we hope to predict the effects of a particular mutation on an exon's splicing efficiency?

Previous Work

The foundations for the approaches in this proposal comes from my postdoctoral work. I had been developing new technologies for gene synthesis using microarray technologies for the purposes of low-cost, high-throughput hypothesis testing. After some initial successes[49,50], it became clear that we would have the ability to generate thousands of synthetic genes, and the question quickly moved to what we could do with these new capabilities. Even if we could generate thousands of hypotheses of how DNA sequences might affect expression, testing all of these reporters individually would require herculean effort. Thus I began to develop rapid ways of quantifying the effects of large reporter libraries using next-generation sequencing.

We designed ~27,000 reporters to test how promoters, ribosome binding sites, and N-terminal codon usage combine to quantitatively determine gene expression levels in *E. coli* (Fig. 2)[51,52]. Briefly, we designed and synthesized the library from DNA microchips and cloned them into a reporter plasmid in-frame with a GFP reporter. After pooled transformation resulting in a bacterial library, we measured DNA and RNA levels using DNASeq and RNASeq. For protein levels, we split the library into 12 equally log-spaced bins, high-throughput sequenced the bins, and used that information to accurately estimate the average protein levels for each construct (FlowSeq). We were able to show quantitative relationships between translation rates and mRNA stabilization as well as N-terminal codon use and translation rates, and resolved long-standing controversies regarding mechanisms for these effects. These experiments were vital for this proposal for two reasons. First,

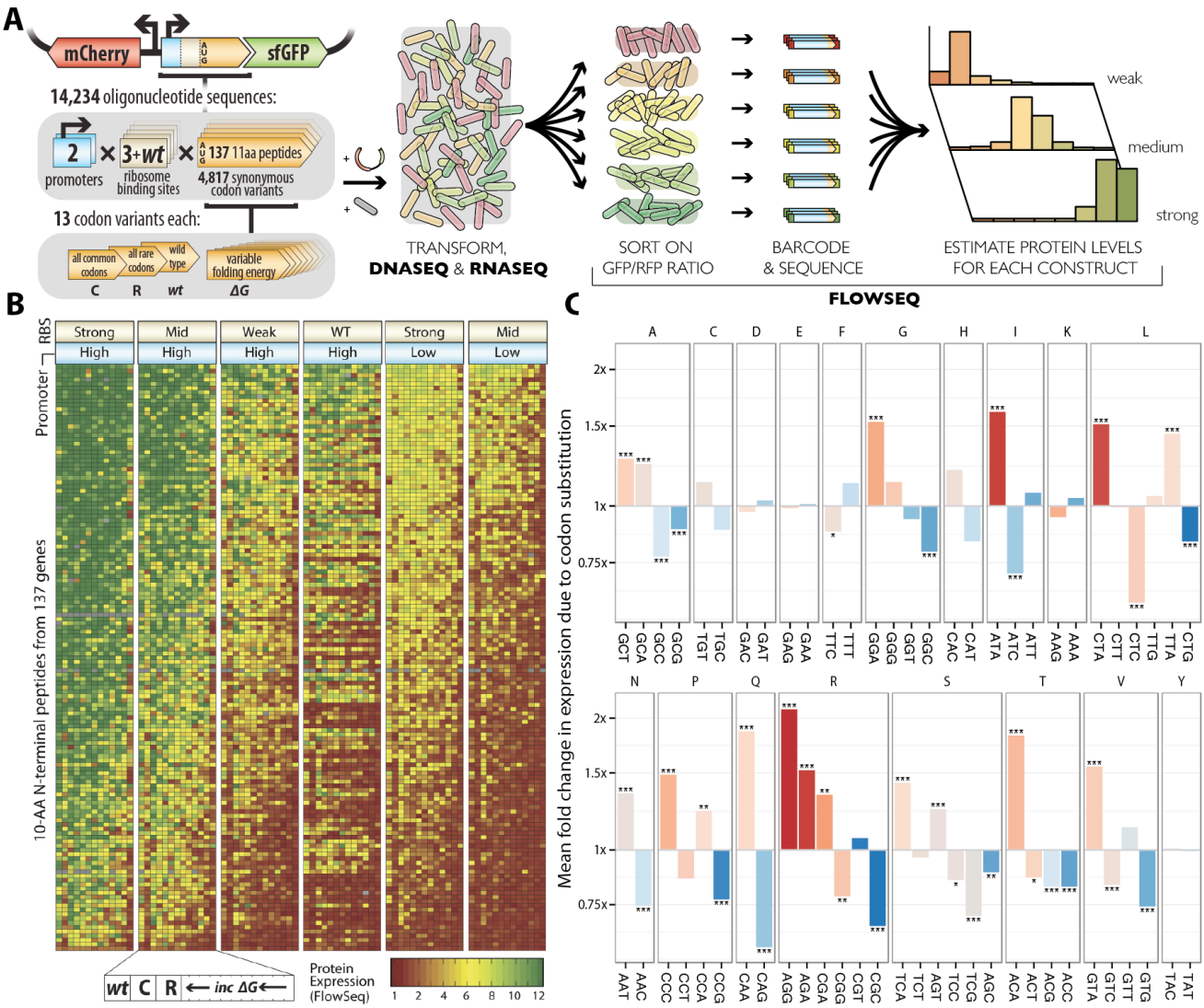

Fig. 2. Multiplexed Analysis of Gene Expression. We have developed methods to measure RNA and protein expression (A) from thousands of designed reporter constructs (>10,000) in a single experiment (B) and used them to answer questions on how regulatory and gene sequences affects quantitative levels of expression (C)[51,52].

they demonstrate that we can accurately quantify RNA and protein expression from large CRE libraries. The techniques including DNASeq, RNASeq, and FlowSeq can easily be used in mammalian cell lines. Second, they indicate that such large multiplexed reporter experiments allow analysis of a large number of dependent processes in order to establish causality and their quantitative importance.

Technology Development

In order to apply these new tools to study regulatory elements governing expression in human cell lines, we will develop a suite of new technologies. First, the modes of regulation in mammalian systems are more complex (e.g., splicing), and will require more classes of multiplexed measurements than required for bacteria (*Multiplexed Assays*). Second, CREs controlling expression in mammalian systems are much longer. Our ability to construct longer synthetic gene libraries will be critical to explore sequence/function relationships in human cell lines (*Gene Library Synthesis)*. Third, our ability to generate large reporter libraries in mammalian cells is more difficult and will require new genome engineering approaches (*Genome Engineering*). Finally, since these CREs interact in complex ways, we will require sophisticated computational analyses for both the design of libraries and the interpretation of the results (*Informatics*).

Multiplexed Assays: In order to study functional genetic elements using multiplexed approaches described here, we require methods to link phenotypic affect on expression with genotype using next-generation sequencing. Our previous developments show that we can link genotype and phenotype for transcriptional and translational processes[51,52]. However, gene regulation in mammalian systems can be much more complex, and and the study of splicing will require a specialized multiplex reporter. Our approach (Fig. 3) uses a novel multiplexed splicing reporter containing a 3 exon, 2 intron gene that codes for Emerald GFP[53], which is connected in-frame by the self-cleaving 2A peptide[54] to mCherry[55]. The middle exon codes for the exon we are interested in examining, and the ratio of GFP to RFP reflects the splicing efficiency of that exon. This reporter construct is based upon other recent screens for splicing regulator elements using FACS[35,56] and is designed to be clonally integrated into human cell lines using site-directed integration into the AAVS1 safe-harbor locus[57] (see *Genome Engineering*). We will use our already established FlowSeq procedures to quantify the library in multiplex. Briefly, we will sort cells into ten bins based on the GFP:mCherry ratio given a minimum mCherry expression level. We will then quantify which constructs fell into each bin using using Illumina paired-end 150bp reads. Since construct sizes will initially be <300bp, this allows for overlapping and thus higher quality reads. Given 600 million paired end reads in a single HiSeq 2500 run, we should have at least 200x coverage for library sizes up to two million individual reporter constructs in a single experiment, or split amongst multiple experiments through barcoding.

A number of features will make this approach useful for these screens. First, unlike previous approaches, we will use two-color sorting. GFP levels are normalized to RFP levels to correct for expression heterogeneity. Second, previous efforts at developing similar systems have suffered because they could not distinguish between transcriptional effects and splicing[58]. Our construct's internal controls will only be measured at the protein level, thus focusing on the effect of splicing. Third, lack of RFP signal can be used as an indicator of single-base deletions, which will cause non-sense mediated decay as well as frameshifting of mCherry. We expect ~10-50% of constructs in our libraries to have a single-base deletion given current error rates of our library construction protocols[50]. Thus we can improve library quality by removing cells that lack RFP signal. Fourth, two TypeIIs restriction sites are designed into the middle of the introns to allow for directional and seamless integration of exons into the reporter construct. Finally, common sequences that allow for robust PCR amplification of the exons after sorting are designed into the common portion of the intron sequence.

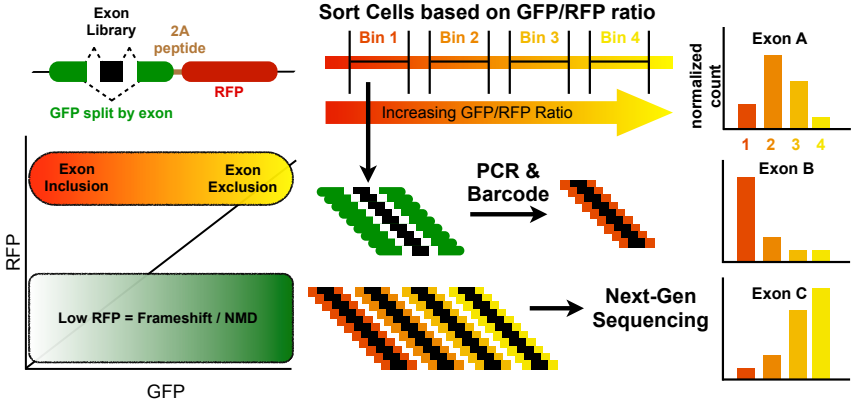

Figure 3. Multiplexed Splicing Reporter. By linking the exon splicing efficiency to the GFP:RFP ratio, we have constructed multiplexed reporters that let us analyze exon splicing efficiencies in multiplex.

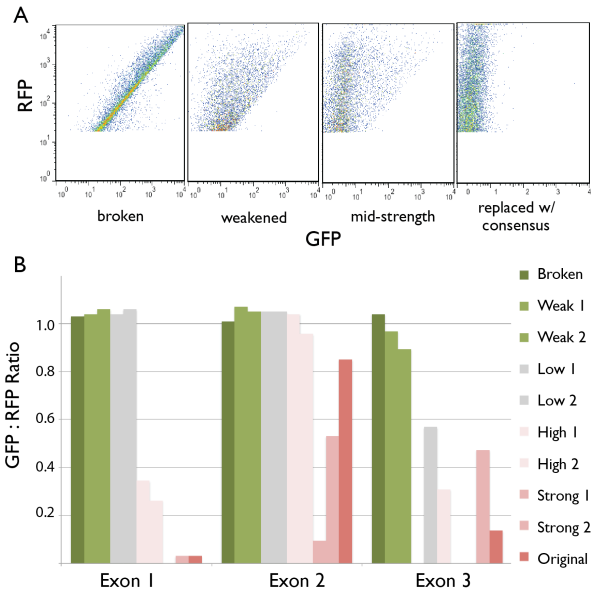

Figure 4. Transient transfections of the multiplexed splicing reporter. Initial data from transient transfections in HEK293 of our multiplexed splicing reporter. We tested mutants of 3 exons and results indicate that when the exon is skipped, you get 1:1 GFP:RFP ratios, and when excluded 1:0 (A), and intermediate levels of splicing are quantitatively reflected in the ratios (B).

Our initial experiments with this splicing reporter show promising results (Fig. 4).

Gene Library Synthesis: Recent advances in gene synthesis are what have made proposals like this one possible[59]. We and others have been applying these tools to study short (<150bp) regulatory enhancer sequences in bacteria, yeast, and mammalian systems[60-64]. The challenge however is that eukaryotic regulatory sequences governing a wide-range of activities are often longer than can be encoded by a single oligonucleotide. Thus methods for constructing longer libraries of sequences are needed in order to resolve outstanding problems in mammalian gene regulation.

For example, as annotated by CCDS, there are 244,211 exons in the human genome spread over 23,754 genes[10]. Of the 190,355 annotated coding exons (exons that do not include 5' and 3' untranslated regions), 97% are less than 300bp in length. In addition to the exon sequence, the intron sequence flanking each exon is necessary to provide accurate measurement of splicing efficiency in exogenous contexts. Previous studies have shown that most intronic splicing regulatory sequences are contained within 40bp of the splice junctions[8,32,33]. Given these constraints, synthesizing a library of 25%, 50%, or 75% of all coding exons with 40bp of intronic sequence on each end would require construct sizes of 170bp, 200bp, and 240bp respectively. However, longer-range interactions are known to affect splicing, and thus the ability to synthesize longer DNA constructs will be useful. More importantly, improving length scales will allow us to use these same methods to construct libraries of other genetic elements such as homotypic clusters of transcription factor binding sites, which average ~600bp[65].

Here, we will utilize two approaches for generating our synthetic gene libraries. First, and most straightforward, we continue to work with most major oligo library providers (Agilent, Gen9, CustomArray) to help develop and improve oligonucleotide libraries using our previously developed next-gen screens for quantifying error rates and distributions. Currently, we have access to oligo libraries of up to one million sequences of up to 300nt long. As mentioned, synthesizing longer gene libraries would be tremendously useful. Thus the major efforts of this proposal are to develop a multiplexed protocol for assembly of long DNA libraries from the microchip using emulsions. As shown in Fig. 5, we have previously developed a parallel approach for gene assembly using oligo pools that we have since commercialized[49]. In this new approach, we can multiplex the synthesis of genes using emulsion-based techniques. We will attempt to construct assemblies of 1000 construct libraries of up to 6 assembled oligos, resulting in constructs up to 730bp in length. At this length, we can synthesize 97% of all exons with up to 200 bases of intronic sequence on each side. Our initial efforts for producing 250-plex assemblies of synthetic genes of up to 500nt long have been successful (*manuscript in preparation*).

TechDev – Genome Engineering: In moving from bacterial to mammalian cell lines, we must overcome several technical hurdles. For the described multiplexed splicing reporter, we require that each cell have only a single reporter construct for quantification using FlowSeq. In addition, clonal integration into the genome at specified locations will allow testing at endogenous expression levels, reduce heterogeneity, and allow us to dissect very particular genomic regulatory positions. Finally, library sizes must be large enough to accommodate our expected synthetic throughput. We will develop methods using targeted Cas9-based nucleases combined with site-specific phage integrases to allow for the construction of high-efficiency, site-directed, clonal reporter library construction in cell lines[66]. For example, in collaboration with Ron Weiss (MIT) we have integrated our splicing reporter into the AAVS1 safe-harbor locus using BxBI integrase to allow for large >1M individual integrations in a lab-scale (T225 Flask) experiment (*data not shown)*. We will expand these efforts to introduce landing pads in new cell lines as well as in new locations (e.g., ROSA26 or specific genomic sites of interest).

This involves integration of an integrase landing-pad that encodes hygromycin resistance. Reporter constructs include a recombinase site that allows integration of a promoter-less puromycin cassette in place of the hygromycin conferring puromycin resistance (and hygromycin sensitivity).

TechDev – Informatics: The reverse genomics approaches we will develop here require new tools for the design and analysis of thousands experiments simultaneously. Rather than extracting information from natural sequences, we must become adept at designing libraries that will best answer questions on how combinations of regulatory elements act in concert to direct the various processes governing splicing. To this end, we will develop informatics tools that will help automate the design of such hypothesis libraries and then analyze and learn from such experiments to design the next iteration. This will require us to compile and annotate all putative regulatory elements found in other studies within exons/introns in the human genome. Then we will develop a framework for mutating and altering these sites to automatically design libraries of exons for testing. We will continue to develop a tool that we have created called SpliceMOD (Fig. 6) that integrates existing datasets for regulatory elements in splicing and combines them into an integrated framework that can annotate genomic sequences, systematically mutate them individually and in groups, and outputs large libraries of exons for testing. During the course of this grant, we will continue to extend this software to 1) integrate new datasets from ENCODE and elsewhere to be more comprehensive, especially with regard to secondary structure, 2) take feedback from analysis of multiplexed reporter assays and define the next set of goals based on machine learning techniques, and 3) modularize and extend the code-base to apply them to regulatory elements controlling other aspects of gene expression.

Pre-mRNA Splicing

We will use these technological developments to dissect CREs controlling pre-mRNA splicing. We will improve the splicing code by constructing exon libraries that will quantitatively test the effects of CREs on splicing. *First*, we will test a library of natural exons to see how well our splicing reporters recapitulate splicing levels of the endogenous exons (*Natural Exons*). *Second*, we will identify known regulatory motifs and features in our natural exon libraries and systematically mutate or move them to test their direct effect on splicing (*Splicing CREs*). *Third*, we will construct a designer exon library that will test a large number of exon/intron pairs coding for the same sequence to understand how such splicing elements act given constraints on amino acid sequence (*CRE Constraints*). *Finally*, we will explore how these CREs act in different cellular contexts by both

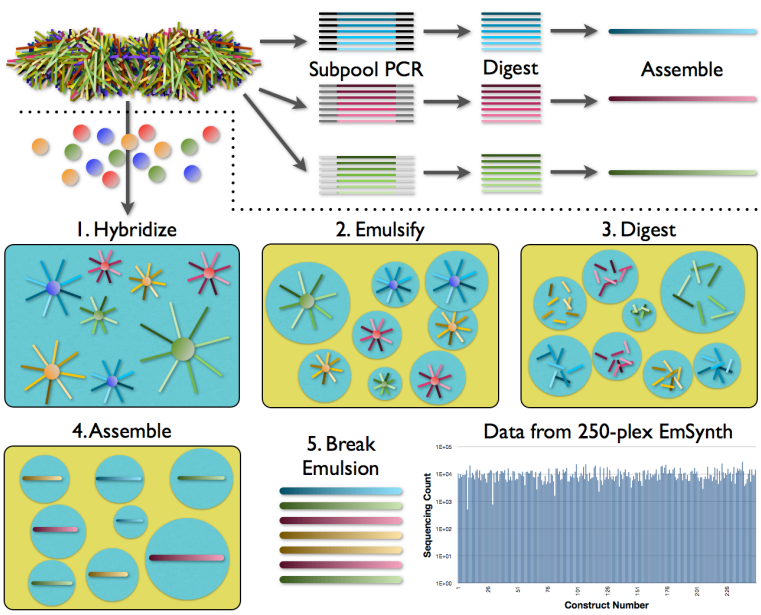

Fig 5. Multiplexed Gene Synthesis. Compared to parallel gene synthesis methods[6,7] (*top*), constructing pools of synthetic genes in multiplex using emulsions will engender cheap, large-scale experiments (*bottom*). Initial data shows synthesis of 250 chorismate mutase variants in a single emulsion PCR.

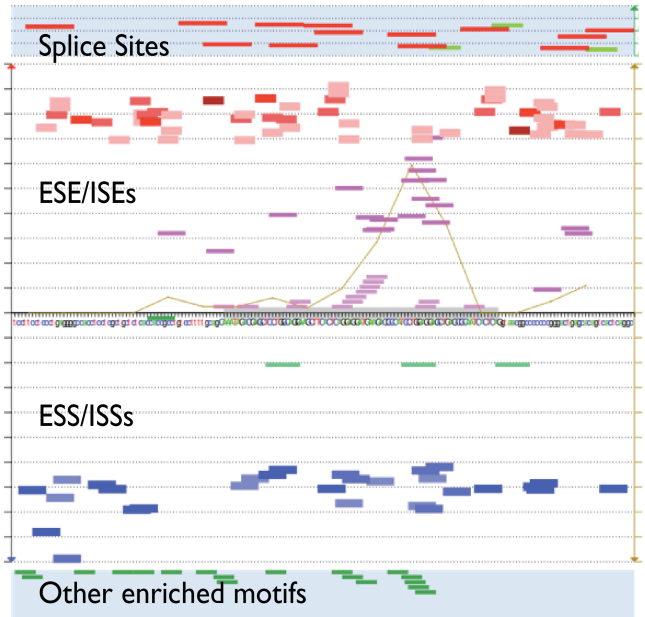

Figure 6. SpliceMod. Known splicing CREs for a human exon are identified by SpliceMod. SpliceMod also generates mutants that modify annotated CREs individually and in combination to test hypotheses using multiplexed reporter libraries.

modulating levels of important splicing regulators *and* testing exon libraries in different human cell lines to better understand alternative splicing (*Alternative Splicing*).

Natural Exons: We will construct libraries of natural exons, clone them into our multiplexed splicing reporter in HEK293, and measure exon splicing efficiencies using FlowSeq. We will use HEK293 as our initial cell line because our integrase landing pad is already integrated into it and most existing splicing CRE studies also use this cell line. We will also perform RNASeq and measure exon inclusion rates of endogenously expressed exons, allowing us to compare exon inclusion rates in our reporter constructs versus their endogenous transcripts. Differences in splicing eficiencies between the reporters and the natural sequences will point to the importance of cis-regulatory regions not included in our initial analyses and will help us understand how much intronic sequence is required to recapitulate endogenous behavior. We have already synthesized a library of 10,000 natural human coding exons that are less than 100bp in length (with an additional 70nt of intronic context) and in-frame with our splicing reporter. We also include polymorphisms found in human populations for a subset of these exons. Based on these pilot studies, we will expand to include more exons using longer oligo libraries and multiplexed gene assemblies to study both longer exons as well as to understand how much longer intronic context contribute to splicing efficiencies.

Splicing CREs: We will use SpliceMod to develop automated methods to construct libraries that test thousands of hypotheses for how CREs affect splicing in our reporter system to refine models of CRE function. For our initial pilot experiment, we used SpliceMod to design 68 mutations for each of 250 randomly chosen natural exons and synthesized the oligo library (17,000 200nt sequences). These mutants will individually and in combination change splicing signals and CREs affecting splicing. Moving forward, we will use the list of 1,014 known or hypothesized splicing regulatory elements as compiled by Barash et al[8], and use it to computationally identify possible regulatory elements in the library of natural exons. Then, we will construct mutants of these sites using different methods depending on whether the elements are located within intronic or exonic sequence. If the regulatory elements are intronic, we will encode single and multiple point mutations as well as move the sequence to different positions on the intron. For exonic regulatory elements, we will use degeneracy in the genetic code to mutate the sequences, while ensuring that new elements are not constructed. In addition, we will test other hypotheses such as the effect of secondary structure on CRE recognition and function. We anticipate that we will conduct many iterations of these experiments that will systematically refine and improve our computational models for how CREs direct splicing.

CRE Constraints: Coding exons have sequence constraints because they must encode a functional protein. Understanding the extent to which CREs within exons can be modulated given an amino acid sequence is important for understanding how such exons can evolve regulatory controls and predicting which changes might have large effects on splicing. We will construct a three-exon, two intron version of our splicing reporter where the middle coding exon encodes a part of the GFP sequence. We will use the degeneracy in the genetic code to generate variants of the middle exon that are predicted to alter splicing efficiencies and quantify the library using FlowSeq. Using SpliceMod, we can computationally enrich or deplete previously identified sequence elements to test their effects on splicing. This library will be cloned into a range of intronic sequences with varying levels of predicted splicing efficiency. These data will also act as an independent test of our splicing models explored above.

Alternative Splicing: The first three efforts all focus on understanding how splicing occurs in a single constant cell type. However, alternative splicing is known to play a large role in differential gene expression in different tissue types[8,42,43]. Splicing differences are thought to occur due to differential regulation of splicing machinery within different cell types. The sequences that direct these alternative splicing effects are thought to occur in *cis*, and have been greatly expanded in human neuronal tissues[42]. However, the CREs that lead to these functional changes have mostly been examined by correlation with tissue-specific RNA-seq. The purpose of this effort is to better understand how the levels and composition of the spliceosomal components affect splicing and how this is manifested in different cell types.

We will first direct efforts at understanding contributions of individual splicing machinery components by modulating individual proteins in the splicing machinery that partake in alternative splicing by siRNA knockdown or transgenic expression of individual proteins. We will construct libraries of natural, mutated, and wholly synthetic exons and measure splicing efficiencies given these perturbations. Candidate proteins that we

will perturb in HEK293 include genes such as U1-5, SR Complex, and other spliceosome components. We will also test these libraries in three ENCODE cell lines: HepG2 (ATCC#:HB-8065), HUVEC (Lonza), and Hela-S3 (ATCC#:CCL-2.2). They represent endoderm, mesoderm, and ectoderm lineages respectively. In addition, we will use two neuronal cell lines, U87 MG (ATCC#:HTB-14) and T98G (ATCC#:CRL-1690), which have previously shown to have neuronal specific gene expression and are easily transformable. For all five cell lines, we will integrate landing pads site-specifically and conduct our splicing efficiency tests on these exon libraries as well as characterize levels of the examined splicing factor levels. Taken together, these experiments will help answer questions on how particular cell types can modulate splicing factor levels to result in particular splice isoforms.

Impact

This proposal outlines methods for large-scale multiplexed synthetic exon experiments to rapidly test thousands of individual hypotheses about exon definition. This work is significant for a number of reasons. First, construction of natural exons with and without mutations of identified regulatory elements will allow direct testing of how important particular elements are to splicing of an exon. Second, differences in levels of splicing for exons in their natural genomic loci and within synthetic constructs will highlight gaps in our understanding for longer-range interactions of cis-elements or other non-sequence specific effects on splicing. Third, testing large numbers of synthetic exons will allow both refinement and validation of how regulatory elements control splicing in cell lines. Finally, studies of alternative splicing usually rely on tissue-specific changes in exon inclusion by averaging potentially hundreds of different cell types. Construction and characterization of natural and synthetic exon libraries in different immortalized cell types will give clearer genome-wide data of how alternative splicing is regulated, and thus how particular mutations affect function.

More broadly, this work will push reverse and functional genetic experiments into the genomics age. The approaches described here will be generally applicable to a wide variety of CREs. For example, understanding how cis-sequences modify methylation and histone patterns are important for understanding establishment and maintenance of epigenetic phenomena. As another example, micro-RNA target sites in 3' UTR affect many transcripts and the rules for how these sequences interact to quantitatively control transcription and translation levels are still unclear. Both of these examples, and many more, are accessible to the reverse genomic technologies and approaches developed in this proposal. Deciphering how our genome sequence regulates function will allow us to better understand how sequence variation affects traits, and how genetic interventions and therapies might be better designed. Finally, the new DNA synthesis technologies described here will lead to progress in not just genetics but in many areas of biology and technology.

B. Innovativeness

Reverse genetics approaches to elucidating how sequence drives function are not particularly new, but they are very powerful. For example, despite intense biochemical and genetic efforts to elucidate the genetic code, it was the first application of nucleic acid synthesis that actually cracked the code[67]. The reason for this is that natural sequence is often not suited to discriminate between possible models and hypotheses, and thus we are left teasing apart models for how sequence defines function[68]. Today, both genetic and biochemical techniques have entered the genomics age brought on by advances in next-generation sequencing. These techniques give us more power than ever at generating and refining hypotheses for how genotype affects phenotype, but testing those hypotheses often still rely on tedious one-at-a-time reporter and genetic modification studies. This proposal tries to flip this paradigm on its head. Here we attempt to scale these tedious one-at-a-time reverse genetic studies to genomic scales. If successful, we will develop powerful next-generation hypothesis testing technologies that complement sequencing-based hypothesis generation capacity. These technologies will not only be applicable to splicing, but a wide-variety of other CREs controlling transcription, translation, localization, methylation, and chromosomal conformations.

While this proposal on its face is risky, we have conducted preliminary experiments showing that the fundamental technological developments are possible. Importantly, we have developed and used a wide-variety of new DNA synthesis technologies that pave the way for these types of experiments. But just as next-generation sequencing alone was insufficient for the progress we see today in genomics, DNA synthesis technologies must be paired with the requisite technologies for design, measurement, and analysis to be useful. To this end, we have already demonstrated that the FlowSeq approach can quantitate ratiometric

fluorescence reporters in multiplex. Additionally, we have shown that the multiplex reporter of splicing works in human cell lines and that the fluorescence ratio correlates with expected levels of splicing. With Ron Weiss (MIT), we have shown that the integrase-based library construction approach can efficiently construct large, clonal libraries at defined positions in human cell lines. Finally, we have built a prototype design and analysis tool, SpliceMod, that lets us build and learn from large reporter libraries to test thousands of hypotheses on splicing CREs in an automated fashion. This proposal seeks to further develop, refine, and integrate these technologies to see if we can get closer to understanding how CREs direct pre-mRNA splicing. More importantly, successes here will help develop a new reverse genomics paradigm to tackle genomic-level questions of how sequence drives gene regulation in humans.

C. Investigator Qualifications

As an undergraduate at UC Berkeley, I majored in the newly formed Bioengineering department and worked in Adam Arkin's lab for three years as one of the first members of the lab. During this time I developed a foundation in computational systems biology and an appreciation for how noisy chemical systems can coordinate very reproducible and interesting behavior. As a graduate student in Biological Engineering at MIT, I joined as the first member of Dr. Drew Endy's laboratory. I worked on understanding the gene expression program of bacteriophage T7. I improved our ability to predict gene expression in T7 by developing more detailed simulations, models, and genome-wide characterizations of gene expression over the course of infection[69]. In addition, I refactored the T7 genome to construct an organism that is a more direct representation of the models that we build. We physically defined, separated, and enabled independent manipulation of primary genetic elements for a large fraction of the wild-type genome and showed that the synthetic refactored phage was viable[70]. This work was the first of now many examples of using genetic refactoring to better understand genetic systems[71-77]. During this time, I with another graduate student in the lab started a wiki for biologists called OpenWetWare (http://openwetware.org), and successfully wrote and received 3 grants to help support hiring and resources for the site. I also helped start MIT's Synthetic Biology Working Group, out of which I helped with several initiatives aimed at growing nascent field such as the starting of the International Genetically Engineered Machines competition and parts registry (http://igem.org). After graduate school, I decided to help start a company (Joule Unlimited) to engineer photosynthetic organisms to make biofuels and joined Flagship Ventures and then Joule as the first employee. In less than two years, we developed plans, set up the company and space, quickly showed many proof-of-concept fuel products, developed precise measurements of carbon flow, and improved flow to our chemical products of interest, which included ethanol and alkanes. As the company grew larger and our initial goals were met (now >100 people with >$150M in total funding), I left to focus on more long-term objectives and joined the Church lab when the Wyss Institute was first started.

At the Wyss, I began working on technologies to increase the scale and reduce the cost of gene synthesis. Both at Joule and at MIT, constructing DNA would be the most limiting step for piloting genetic designs. Technologies to reduce the cost an order-of-magnitude would increase both the time and scope researchers have to conduct experiments. I developed reliable and scalable protocols to do just that, and transitioned this work in 2011 to a startup company called Gen9, which is now the commercial leader for cheap synthetic genes to industrial and academic labs. In turn, I transitioned my focus to leveraging cheap DNA to ask questions in biology that have been difficult to tackle otherwise by developing multiplexed measurements of transcription and translation of large reporter libraries using next generation sequencing. These works are the basis of this proposal, and have led to nine papers including ones in *PNAS*[52]*, Nature Biotechnology*[49,66,78], and *Science*[50,51,79]. In addition, I wrote several grants (including an ONR grant specifically for my projects) and mentored two graduate students and a technician.

As shown in this proposal, I spent much of my remaining time at Harvard piloting the technologies that I will require to transition my work to human genomics. Over the last two years, I've trained in cell culture techniques, helped with several human genome engineering projects[66,78], and piloted new emulsion-based gene assembly methods. In addition, as I transition to UCLA, I have overseen lab renovations, developed collaborations that will help with much of this work, already hired my first postdoc, recruited several graduate students to begin rotations as I arrive, and put in place collaborative frameworks with the industry leading oligo library providers (Agilent, CustomArray, Gen9).

I am determined to use my training in technology development and bacterial gene expression to tackle problems in human biology and disease. Though my career has not been the most linear of paths, I have developed deeper experiences in computation, engineering, molecular biology, genetics, gene regulation, bioinformatics, next-generation sequencing, systems biology, synthetic biology, manuscript and grant writing, entrepreneurship, mentorship and management, and new lab setup than most beginning PIs. I feel that these experiences will serve me well as I begin a new lab in January 2014. The New Innovator Award will give me the freedom I need to really develop and integrate the technologies described here and use them to engender powerful new methodologies to study human genetics. Once established, I am confident that I can transition this work to more stable, hypothesis-driven funding mechanisms.

REFERENCES CITED

1. Venter, JC, et al., *Science*, (2001). **291**:1304
2. Lander, ES, et al., *Nature*, (2001). **409**:860
3. Collins, F, *Nature*, (2010). **464**:674
4. Venter, JC, *Nature*, (2010). **464**:676
5. Consortium, EP, *Science*, (2004). **306**:636
6. Consortium, EP, et al., *Nature*, (2012). **489**:57
7. Consortium, EP, et al., *Nature*, (2007). **447**:799
8. Barash, Y, et al., *Nature*, (2010). **465**:53
9. Matlin, AJ, F Clark, and CW Smith, *Nat Rev Mol Cell Biol*, (2005). **6**:386
10. Pruitt, KD, et al., *Genome Res*, (2009). **19**:1316
11. Keren, H, G Lev-Maor, and G Ast, *Nat Rev Genet*, (2010). **11**:345
12. Matlin, AJ and MJ Moore, *Adv Exp Med Biol*, (2007). **623**:14
13. Jurica, MS and MJ Moore, *Mol Cell*, (2003). **12**:5
14. Faustino, NA and TA Cooper, *Genes Dev*, (2003). **17**:419
15. Poulos, MG, et al., *Cold Spring Harb Perspect Biol*, (2011). **3**:a000778
16. Tazi, J, N Bakkour, and S Stamm, *Biochim Biophys Acta*, (2009). **1792**:14
17. Pan, Q, et al., *Nat Genet*, (2008). **40**:1413
18. Wang, ET, et al., *Cell*, (2012). **150**:710
19. Robberson, BL, GJ Cote, and SM Berget, *Mol Cell Biol*, (1990). **10**:84
20. Berget, SM, *J Biol Chem*, (1995). **270**:2411
21. Carothers, AM, et al., *Mol Cell Biol*, (1993). **13**:5085
22. Krawczak, M, J Reiss, and DN Cooper, *Hum Genet*, (1992). **90**:41
23. Schwartz, SH, et al., *Genome Res*, (2008). **18**:88
24. Senapathy, P, MB Shapiro, and NL Harris, *Methods Enzymol*, (1990). **183**:252
25. Zhang, XH, CS Leslie, and LA Chasin, *Methods*, (2005). **37**:292
26. Kol, G, G Lev-Maor, and G Ast, *Hum Mol Genet*, (2005). **14**:1559
27. Smith, CW and B Nadal-Ginard, *Cell*, (1989). **56**:749
28. Green, MR, *Annu Rev Cell Biol*, (1991). **7**:559
29. Sun, H and LA Chasin, *Mol Cell Biol*, (2000). **20**:6414
30. Trifonov, EN, *Comput Appl Biosci*, (1996). **12**:423
31. Fu, XD, *Cell*, (2004). **119**:736
32. Chasin, LA, *Adv Exp Med Biol*, (2007). **623**:85
33. Wang, Z and CB Burge, *RNA*, (2008). **14**:802
34. Arias, MA, S Ke, and LA Chasin, *Nat Biotechnol*, (2010). **28**:686
35. Wang, Z, et al., *Cell*, (2004). **119**:831
36. Fairbrother, WG, et al., *Science*, (2002). **297**:1007
37. Smith, PJ, et al., *Hum Mol Genet*, (2006). **15**:2490
38. Zhang, XH and LA Chasin, *Genes Dev*, (2004). **18**:1241
39. Friedman, BA, et al., *Genome Res*, (2008). **18**:1643
40. Ke, S and LA Chasin, *Genome Biol*, (2010). **11**:R84
41. Barash, Y, BJ Blencowe, and BJ Frey, *Bioinformatics*, (2010). **26**:i325
42. Barbosa-Morais, NL, et al., *Science*, (2012). **338**:1587
43. Merkin, J, et al., *Science*, (2012). **338**:1593
44. Genomes Project, C, et al., *Nature*, (2010). **467**:1061
45. Rios, J, et al., *Hum Mol Genet*, (2010). **19**:4313
46. Roach, JC, et al., *Science*, (2010). **328**:636
47. Richard, P, et al., *Neuromuscul Disord*, (2007). **17**:409
48. Pagani, F, M Raponi, and FE Baralle, *Proc Natl Acad Sci U S A*, (2005). **102**:6368
49. Kosuri, S, et al., *Nat Biotechnol*, (2010). **28**:1295
50. Church, GM, Y Gao, and S Kosuri, *Science*, (2012). **337**:1628
51. Goodman, DB, GM Church, and S Kosuri, *Science*, (2013). 10.1126/science.1241934
52. Kosuri, S, et al., *Proc Natl Acad Sci U S A*, (2013). **110**:14024
53. Tsien, RY, *Annu Rev Biochem*, (1998). **67**:509
54. Szymczak, AL, et al., *Nat Biotechnol*, (2004). **22**:589
55. Shaner, NC, et al., *Nat Biotechnol*, (2004). **22**:1567
56. Culler, SJ, et al., *Nucleic Acids Res*, (2010). **38**:5152
57. DeKelver, RC, et al., *Genome Res*, (2010). **20**:1133
58. Ke, S, et al., *Genome Res*, (2011). **21**:1360
59. LeProust, EM, et al., *Nucleic Acids Res*, (2010). **38**:2522
60. Smith, RP, et al., *Nat Genet*, (2013). **45**:1021
61. Patwardhan, RP, et al., *Nat Biotechnol*, (2012). **30**:265
62. Melnikov, A, et al., *Nat Biotechnol*, (2012). **30**:271
63. Kheradpour, P, et al., *Genome Res*, (2013). **23**:800
64. Sharon, E, et al., *Nat Biotechnol*, (2012). **30**:521
65. Gotea, V, et al., *Genome Res*, (2010). **20**:565
66. Mali, P, et al., *Nat Biotechnol*, (2013). **31**:833
67. Nirenberg, MW and JH Matthaei, *Proc Natl Acad Sci U S A*, (1961). **47**:1588
68. Brenner, S, *Philos Trans R Soc Lond B Biol Sci*, (2010). **365**:207
69. Kosuri, S, JR Kelly, and D Endy, *BMC Bioinformatics*, (2007). **8**:480
70. Chan, LY, S Kosuri, and D Endy, *Mol Syst Biol*, (2005). **1**:2005 0018
71. Fitzgerald, JT, et al., *J Am Chem Soc*, (2013). **135**:3752
72. Ghosh, D, et al., *ACS Synth Biol*, (2012). **1**:576
73. Lim, JH, et al., *Bioresour Technol*, (2013). **135**:568
74. Nano, FE, *Curr Opin Biotechnol*, (2012). **23**:897
75. Shao, Z, et al., *ACS Synth Biol*, (2013). 10.1021/sb400058n
76. Springman, R, et al., *ACS Synth Biol*, (2012). **1**:425
77. Temme, K, D Zhao, and CA Voigt, *Proc Natl Acad Sci U S A*, (2012). **109**:7085
78. Zhang, F, et al., *Nat Biotechnol*, (2011). **29**:149
79. Lajoie, MJ, et al., *Science*, (2013). **342**:357

FACILITIES & OTHER RESOURCES

Note: PI arrives at UCLA in January 2014. Laboratory/Office space and basic equipment will be ready upon his arrival.

Laboratory/Office: The PI's wet and computational laboratories and office space occupy approximately ~1500 sq. ft. on the sixth floor of Boyer Hall and is equipped for BL1 and BL2 cell culture, micro/molecular biology, and microscopy. UCLA has provide the PI with $1.3 million in unrestricted funds to outfit the laboratory in addition to separate funds for laboratory renovation, furniture, core facility access, and recurring endowed chair income.

NGS Sequencing: The PI will have access to four core facilities at UCLA including the Broad Center for Regenerative Medicine, Pathology, Human Genetics, and Neurosciences that include services and equipment for NGS on Roche and Illumina platforms including HiSeq 2000/2500, Miseq, 454, and Ion Torrent PGM. In addition, the PI often outsources NGS sequencing to Beijing Genomics Institute and has access to a PacBio RS instrument through collaboration.

Oligo Libraries & Gene Synthesis: The PI has access to both production and research-grade oligo libraries under collaborative agreement with Agilent Technologies. In addition, the PI actively collaborates with CustomArray, another provider of oligo libraries, and has access to a CustomArray microarray synthesizer in collaboration with the Wyss Institute (Harvard). Finally, the PI has special access and pricing to gene library products from Gen9 as a member of the Scientific Advisory Board.

Flow Cytometry and Sorting: The PI will have access core facilities at the Broad Center for Regenerative Medicine and the Johnson Comprehensive Cancer Center at UCLA which combine to include 9 flow cytometers and 6 flow sorters (all Becton Dickenson models).

High Performance Computing: The PI will have access to shared computational infrastructure as a member of the DOE-UCLA Genomics and Proteomics Center, including a computer network consisting of some 35 quad core I7 based workstations. A 200 Terabyte RAID system and a 90 Terabyte tape library is used for continuous backup. Larger computing jobs are performed on a shared 180 node ROCKS cluster. Finally, as a member of the Bioinformatics program faculty, the PI will have access to the 11,000 process Hoffman2 Cluster.

Biochemical Instrumentation Facility and Molecular Instrumentation Center: The PI has free access (until 2020) to these facility, housed in the building adjacent to Boyer Hall, containing state-of-the-art instrumentation for Biochemistry. Potentially relevant equipments includes cell counters (Nexcelom), NanoDrop, UV/VIS Spectrophotometer (HP), Fluor/Phosphoimager (BioRad), Real-time PCR (BioRad), Fluorescense Gel Imager (Alpha Innotech), ultracentrifuge (Beckman Optima XL-A). Other equipment includes many instruments for mass spectrometry (Quad-TOF, MALDI-TOF, GCMS, LCMS/ESI), NMR, 2D Electrophoresis, NRM, TEM, differerential scanning calorimetry

Scientific Environment: At UCLA, all sciences, engineering, and medical departments are physically located close to one another leading to a dynamic and collaborative environment. Already, I have begun interacting/collaborating with colleagues that will be useful for this proposal. Their expertise spans genomics (Leonid Kruglyak, Matteo Pelligrini, Steve Jacobsen, Jason Ernst) splicing (Doug Black), systems biology (Alex Hoffman), stem cell biology (Kathrin Plath).

BIOGRAPHICAL SKETCH

Provide the following information for the Senior/key personnel and other significant contributors in the order listed on Form Page 2.
Follow this format for each person. **DO NOT EXCEED FOUR PAGES.**

NAME	POSITION TITLE
Sriram Kosuri	Assistant Professor Dept. of Chemistry and Biochemistry University of California, Los Angeles
eRA COMMONS USER NAME (credential, e.g., agency login) SKOSURI	

EDUCATION/TRAINING *(Begin with baccalaureate or other initial professional education, such as nursing, include postdoctoral training and residency training if applicable.)*

INSTITUTION AND LOCATION	DEGREE *(if applicable)*	MM/YY	FIELD OF STUDY
University of California, Berkeley (USA)	B.S.	05/01	Bioengineering
Massachusetts Institute of Technology (USA)	Sc.D.	02/07	Biological Engineering
Harvard Medical School (MA, USA)	PostDoc	12/11	Dept. of Genetics

B. Positions and Honors

Positions and Employment

2007-2009 Senior Scientist, Joule Unlimited Inc., Cambridge, MA
2009-2011 Postdoctoral Fellow, Dept. of Genetics & Wyss Institute, Harvard Medical School, Boston, MA
2012-2013 Staff Scientist, Advanced Technology Team, Wyss Institute at Harvard University, Boston, MA
2014- Assistant Professor, Dept. of Chemistry and Biochemistry, UCLA, Los Angeles, CA

Other Experience and Professional Memberships

2003-2009 Founder and Board Member, OpenWetWare.Org, Cambridge, MA
2012- Scientific Advisory Board, Gen9 Inc., Cambridge, MA
2012- Ad Hoc Reviewer for *Nature Methods, Nucleic Acids Research, ACS Synthetic Biology*

Honors

2000 HHMI Summer Undergraduate Fellowship, UC Berkeley, Berkeley, CA
2001 Andrew and Edna Viterbi Fellowship in Computational Biology, MIT, Cambridge, MA
2002 NIH Biotechnology Training Program Fellowship, MIT, Cambridge, MA
2005 Center for Bits and Atoms Training Program, MIT, Cambridge, MA
2012 PopTech Science Science Fellow
2014 Linda and Fred Wudl Term Chair

C. Selected Peer-reviewed Publications

* denotes equal contribution; underline indicates corresponding author.

1. Chan LY*, Kosuri S*, Endy D. Refactoring bacteriophage T7. Molecular Systems Biology. 1:2005 0018. 2005; doi: 10.1038/msb4100025. PubMed PMID: 16729053; PubMed Central PMCID: PMC1681472.
2. Kosuri S, Kelly JR, Endy D. TABASCO: A single molecule, base-pair resolved gene expression simulator. BMC Bioinformatics. 8:480. 2007; doi: 10.1186/1471-2105-8-480. PubMed PMID: 18093293; PubMed Central PMCID: PMC2242808.
3. Kosuri S*, Eroshenko N*, Leproust EM, Super M, Way J, Li JB, et al. Scalable gene synthesis by selective amplification of DNA pools from high-fidelity microchips. Nature Biotechnology. 28(12):1295-9. 2010; doi: 10.1038/nbt.1716. PubMed PMID: 21113165; PubMed Central PMCID: PMC3139991.
4. Zhang F*, Cong L*, Lodato S, Kosuri S, Church GM, Arlotta P. Efficient construction of sequence-specific TAL effectors for modulating mammalian transcription. Nature Biotechnology. 29(2):149-53. 2011; doi: 10.1038/nbt.1775. PubMed PMID: 21248753; PubMed Central PMCID: PMC3084533.

5. Eroshenko N*, Kosuri S*, Marblestone AH, Conway N and Church GM. Gene Assembly from Chip-Synthesized Oligonucleotides. Current Protocols in Chemical Biology. 4:1–17. 2012.
6. Church GM, Gao Y, Kosuri S. Next-generation digital information storage in DNA. Science. 337(6102):1628. 2012; doi: 10.1126/science.1226355. PubMed PMID: 22903519.
7. Mali P, Aach J, Stranges PB, Esvelt KM, Moosburner M, Kosuri S, et al. CAS9 transcriptional activators for target specificity screening and paired nickases for cooperative genome engineering. Nature Biotechnology. 31(9):833-8. 2013; doi: 10.1038/nbt.2675. PubMed PMID: 23907171.
8. Kosuri S*, Goodman DB*, Cambray G, Mutalik VK, Gao Y, Arkin AP, et al. Composability of regulatory sequences controlling transcription and translation in Escherichia coli. Proceedings of the National Academy of Sciences of the United States of America. 110(34):14024-9. 2013; doi: 10.1073/pnas.1301301110. PubMed PMID: 23924614; PubMed Central PMCID: PMC3752251.
9. Lajoie MJ*, Kosuri S*, Mosberg JA, Gregg CJ, Zhang D, Church GM. Probing the Limits of Genetic Recoding in Essential Genes. Science. 342(6156):361-3. 2013; doi: 10.1126/science.1241460.
10. Goodman DB, Church GM, Kosuri S. Causes and Effects of N-Terminal Codon Bias in Bacterial Genes. Science. 2013. doi: 10.1126/science.1241934. PubMed PMID: 24072823.