PROJECT TITLE: Identification of translational regulatory elements

AIMS AND BACKGROUND

When, where, and how much of a protein is expressed is important to explain the underlying biological processes, for example, in the context of developmental biology, or disease etiology. A major determinant of protein expression level is the rate of protein translation. However, unlike other steps in the central dogma, such as transcriptional regulation, much less is known about translational regulatory mechanisms. A particularly intriguing one involves *upstream Open Reading Frames* (uORFs). uORFs are pairs of start and stop codons located upstream of protein coding regions (Figure 1), and translation at uORFs has been known to affect translation of coding regions during important biological processes [6-8], such as hypoxia condition [13] and stress response [14]. However, despite their importance, exactly how uORFs regulate translation remains unclear, and systemic investigations of uORF regulatory function have been challenging due to the limited knowledge of translated uORFs. **This project will develop the methods that enable comprehensive identification of translated uORFs in tissues/organisms of interests, and annotate the most complete sets of translated uORFs in human lymphoblastoid cell lines, which will facilitate our understanding of translational regulatory mechanisms.**

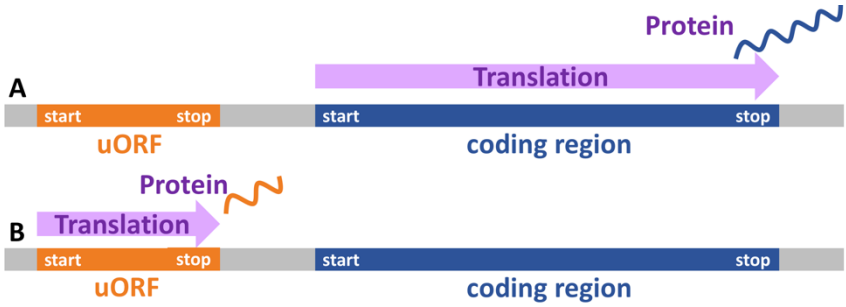

Figure 1 *Example of uORF regulatory mechanism. (A) The translation process bypasses the start codon in the uORF, recognizing the downstream start codon and translating the coding region. (B) The process recognizes and translates the uORF, synthesizing a small protein. If translation termination of the uORF is efficient, the coding region is not translated.*

Recently, ribosome profiling, a technique for directly quantifying levels of translation [2], has enabled genome-wide identification of translated coding regions [1,3-5]. Previous analyses [3-5] of the ribosome profiling data have identified translated uORFs, but they assumed that the translations of uORFs start/stop at only canonical start/stop codons. As co-leading authors, CI Shim and PI Wang successfully identified and experimentally validated novel translated uORFs starting at any start codons in human lymphoblastoid cell lines by modelling the fine-scale structure in the data [1], and observed that ~60% of the uORFs start at non-canonical start codons. Our pioneering work is a promising start to comprehensive identification of translated uORFs. However, there remains considerable room for innovation and improvement that is essential to identify a complete set of translated uORFs. Indeed, the methods we used in [1] were originally designed to identify translated coding regions. Thus, we searched for translated uORFs by restricting our analysis to upstream of translated coding regions which made up less than 10% of the total 431K expressed transcripts. In other words, more than 90% of the transcriptome were not explored for signatures of translated uORFs. Moreover, our analysis probably missed uORFs that completely silence downstream coding regions (Figure 1B).

Therefore, building on our previous analysis in [1], our multidisciplinary investigator team in the current proposal will tackle the following aims:

Aim 1: **Develop novel statistical methods** that leverage the fine-scale structure in ribosome profiling data **to enable the search for translated uORFs over the full transcriptome** using human lymphoblastoid cell lines as an initial model system.
Aim 2: **Experimentally validate the translated uORFs identified in Aim 1** using proteomics and next generation sequencing.

With these aims, this project will provide the most comprehensive annotation of translated uORFs in human lymphoblastoid cell lines, and freely-available software for end-users to identify translated uORFs in their tissues/organisms of interests. By providing these resources, we will facilitate biological insights into regulatory function of uORFs, and add further knowledge to the research field of translation regulation, and ultimately gene regulation.

INVESTIGATOR(S)

To ensure the success of this project, we have assembled a multidisciplinary team consisting of three faculty members with expertise in statistics, computational biology, gene regulation, and proteomics, and two trainees.

CI Shim (20% FTE years 1-3) will provide leadership for the project, conduct administration, and take responsibility for overall management and coordination of the project. As an expert in statistical modelling of fine-scale structures in high-throughput sequencing data, analysis of ribosome profiling data and software development, she will lead the development of the methods, data analysis, software development (Aim 1) and validation using harringtonine-treated ribosome profiling data (Aim 2b). She will provide primary supervision and training for the research associate and PhD student. She has successfully co-supervised to completion one PhD student (PhD passed 2017) and co-mentored one research fellow. She is currently co-supervising four PhD students (one as a main supervisor and three as a co-supervisor).

PI Wang (10% FTE years 1-3) will be based in the University of Texas Health Science Center at Houston in US for the duration of the project, and bring in expertise in translational regulation, ribosome profiling, and proteomics. He will lead the validation step using tandem mass spectrometry in Aim 2a. He will design the proposed tandem mass spectrometry experiments, and provide input in Aim 1 and Aim 2b. He will additionally provide informal mentoring for the research associate and PhD student.

PI Pique-Regi (5% FTE years 1-3) will be based in Wayne State University in US for the duration of the project and chiefly contribute his extensive expertise in computational genomics and bioinformatics to this project. As an expert in signal detection in ChIP-seq and ATAC-seq data, he will lead the identification of translation initiation peak in Aim 2b. He will also provide overall guidance to the PhD student on building a web-based platform (Aim 1c). His group already built multiple web-based platforms such as CentiSNPs and GxE browser. Moreover, he developed a novel method, called CENTIPEDE to identify tissue-specific regulatory elements for DNA-binding proteins using functional genomics data. He will contribute this experience to the development of methods (Aim 1a).

A research associate (100% FTE years 1-3) will be hired and participate in all parts of the project. The research associate will focus on the bioinformatics aspect of the methods development, perform data analysis, implement the methods in user-friendly software (Aim 1), and conduct validations (Aim 2). The role of the research associate will require strong skills in statistical genomics, bioinformatics, computational biology, and programming.

One PhD student (100% FTE years 1-3) will be appointed to work on this project. The PhD student will focus on statistical modelling in the method development (Aim 1a), perform data analysis (Aim 1b), and build tools for effective sharing and visualization of results (Aim 1c). The role of the PhD student will require strong skills in statistics, bioinformatics, and computational biology.

The CI and PIs will have very close communication by having weekly regular meetings via skype. The CI and PIs have done it this way for the previous and other ongoing projects to integrate the expertise of the CI and PIs in every step of the proposed project.

The research associate and PhD student will be based in CI Shim's group at the Melbourne Integrative Genomics (MIG) and the School of Mathematics and Statistics. The trainees will have daily interactions with CI Shim's group members and weekly interactions with PI Wang's group via skype. This project will train at least two early career researchers in Australia to use an emerging data type, ribosome profiling data, for the study of translational regulation.

Our team is capable of building international linkages: Before relocation to Australia in 2017, CI Shim was based in US (PhD, postdoc, and tenure track Assistant Professor). As shown in the published papers, she has worked with leading researchers in US and Europe (e.g., Prof. Matthew Stephens and Prof. Yoav Gilad in University of Chicago, Prof. Jonathan Pritchard in Stanford University, Dr. Eileen Furlong in European Molecular Biology Laboratory in Heidelberg) and she maintains strong collaborative ties with members from the groups of those leading researchers. Indeed, PI Wang is from Prof. Yoav Gilad lab and PI Pique-Regi is from Prof. Jonathan Pritchard lab. PI Wang is based in US and has worked with leading researchers. He is an expert in translational regulation and published two major papers in the filed as a co-first author: Battle A*, Khan Z*, Wang SH* et al, 2014, Science (199 citations) and Raj A*, Wang SH*, Shim H* et al, 2016, eLife (51 citations). PI Pique-Regi is also based in US and has extensive international collaboration networks, as

shown in his publication records. Our team will ensure international uptake and impact of outcomes from this proposal.

Existing collaborations are already in place: CI Shim, PI Wang, and PI Pique-Regi worked as postdoctoral scholars in the department of Human Genetics at the University of Chicago. CI Shim and PI Wang published the eLife paper as co-leading authors which forms the basis of the current proposal. CI Shim and PI Pique-Reqi have been collaborating since 2014 and submitted a paper as co-senior authors (available on bioRxiv).

PROPOSED PROJECT QUALITY AND INNOVATION

Significance of the proposal

Understanding how gene expression is regulated is fundamental to modern molecular biology research. In many cases, when, where and how much of a protein is made is important to explain the underlying biology. This is particularly true in the context of developmental biology, or disease etiology. During the progression of these biological processes, the same genome is expressed differently, which often leads to drastically different phenotypic outcomes.

Although for most biological processes protein expressions are relevant quantities, much of the research on gene regulation has been focusing on transcription. Though informative, transcript levels don't always reflect the level of protein expression. In fact, ample examples in the literature have shown a discordance between transcript and protein levels [9-12]. Therefore, in order to gain a full picture, it is important to have a clear understanding of how transcript level variations are propagated downstream to the protein level.

One major determinant of protein expression level is the rate of protein translation. While the mechanism of translation has been studied in detail using biochemical and molecular approaches, much less is known about its regulation. In contrast to our knowledge of transcriptional regulation and regulatory elements such as promoters and enhancers, much less is known about translational regulatory elements and corresponding mechanisms controlling translation.

A particularly intriguing mode of translational regulation involves a unique type of regulatory element, uORF (Figure 1). Translation at uORFs has been known to affect translation of downstream coding regions during important biological processes, such as hypoxia condition [13] and stress response [14]. However, despite their importance, exactly how uORFs regulate translation remains unclear, and systemic investigations of uORF regulatory function have been challenging due to the limited knowledge of translated uORFs.

Some studies [3-5] used the ribosome profiling data to identify translated uORFs, but their analyses assumed that the translations of uORFs start/stop at only canonical start/stop codons. The method proposed in our eLife paper [1] enabled identification of translated uORFs starting at any start codons, and we observed that ~60% of the uORFs start at non-canonical start codons in human lymphoblastoid cell lines. However, the current method limited our search for translated uORFs to upstream regions of translated coding regions that were only less 10% of the transcriptome in human lymphoblastoid cell lines.

In this project, we will **develop the novel methods that enable more comprehensive identification of translated uORFs** in tissues/organisms of interests, and provide **the most complete annotation of translated uORFs** in human lymphoblastoid cell lines in order to **facilitate new insights into regulatory mechanism of protein translation,** and, ultimately, gene regulation.

Preliminary Studies

Our team has ample expertise in 1) the modelling and analysis of ribosome profiling data in order to identify translated regions, and 2) mass spectrometry and ribosome profiling experiments for validation of translation at those regions: Ribosome profiling is a novel technique used to quantify levels of translation through high throughput sequencing of ribosome-protected RNA fragments [2]. When aggregated across annotated coding transcripts centered at their translation initiation (or termination) sites, ribosome profiling data show two distinct features that reflect mechanisms of translation and distinguish translated regions from untranslated regions (Figure 2A).

- Higher abundance within the translated region: ribosome profiling counts are highly enriched within the translated region. Moreover, base positions within the translated region close to the translation initiation and termination sites have substantially higher ribosome profiling counts compared to base positions in the rest of the translated region.

- Three-base periodicity within the translated region: ribosome profiling counts typically peak at the first position of each codon. The count over the start and stop codons tend to have a stronger peak (thus, a slightly different periodic pattern) compared to the rest of the translated region. The counts in the untranslated regions lack this periodic pattern with similar aggregate counts among base positions.

In our eLife paper [1], we developed a framework to identify the translated coding region in a transcript using a mixture of hidden Markov models (HMMs) that 1) captures these distinct features of ribosome profiling data and 2) integrates RNA sequence information and transcript expression. Using our method, we identified 7,273 novel translated coding regions in human lymphoblastoid cell lines (LCLs). We observed that some of them start at non-canonical start codons (Figure 2B). Moreover, we experimentally validated them using mass spectrometry and harringtonine-treated ribosome profiling data.

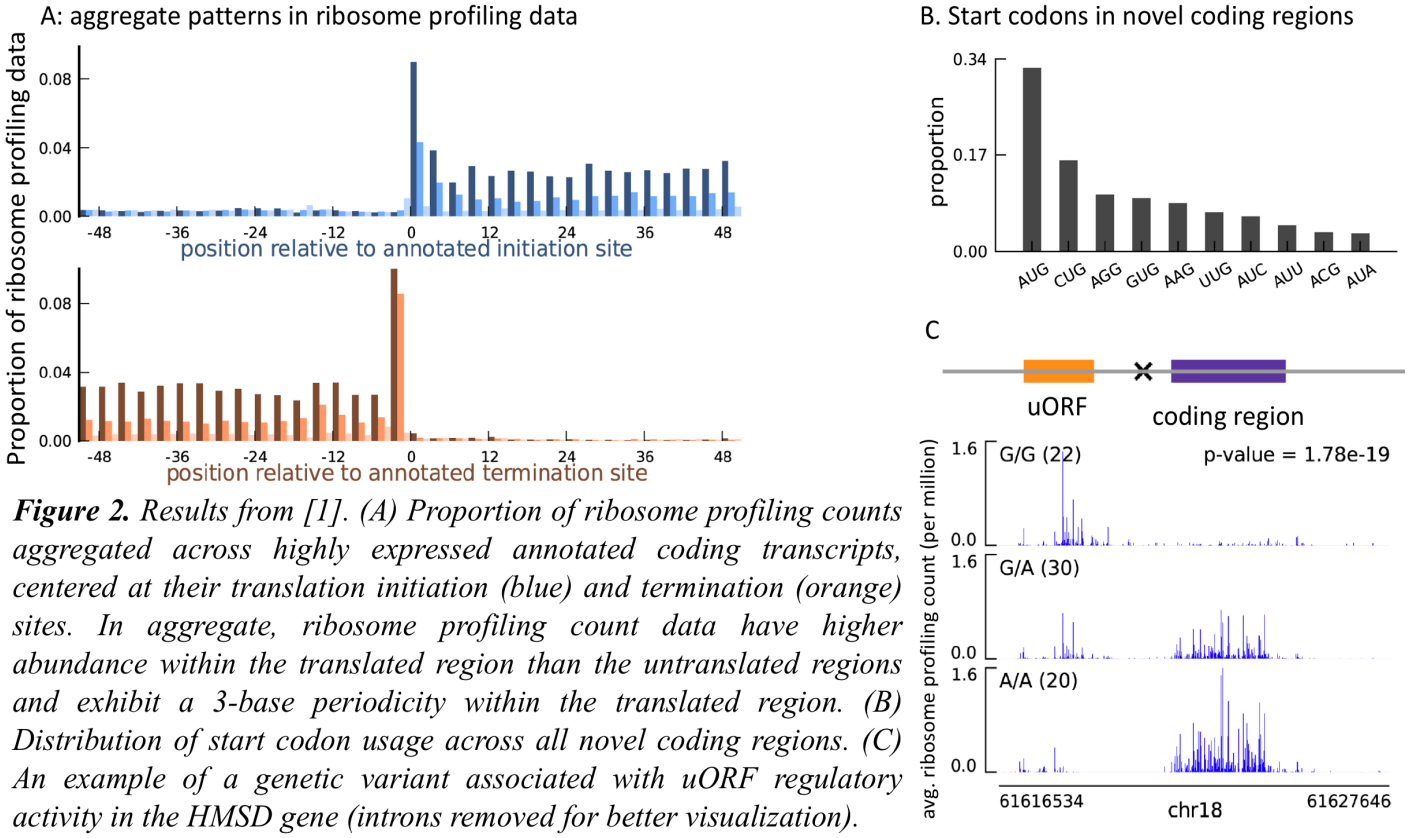

Figure 2. *Results from [1]. (A) Proportion of ribosome profiling counts aggregated across highly expressed annotated coding transcripts, centered at their translation initiation (blue) and termination (orange) sites. In aggregate, ribosome profiling count data have higher abundance within the translated region than the untranslated regions and exhibit a 3-base periodicity within the translated region. (B) Distribution of start codon usage across all novel coding regions. (C) An example of a genetic variant associated with uORF regulatory activity in the HMSD gene (introns removed for better visualization).*

Our team made a promising start to the future studies of uORFs as regulatory elements: Among other findings in our eLife paper [1], we annotated 2,442 translated uORFs. Moreover, we demonstrated that ~40% of these translated uORFs are likely to regulate the translation of their downstream coding regions, as the uORFs and downstream coding regions show negative correlation in levels of translation. Even more interestingly, we identified, for the first time, genetic variants that are associated with uORF regulatory activity (Figure 2C). Our pioneering work is a promising start to the future studies of uORFs as translational regulatory elements. However, there remains considerable room for innovation and improvement that is essential to identify a complete set of translated uORFs. The method we used to identify the translated uORFs in [1] was originally designed to identify translated coding regions. Thus, it allowed us to search for translated uORFs only in upstream regions of 36K translated coding regions which are less than 10% of the total 431K expressed transcripts in human LCLs. In other words, more than 90% of the transcriptome were not explored for signatures of translated uORFs. Moreover, the method possibly failed to identify particular types of uORFs such as those that completely silence downstream coding regions (Figure 1B). Building on our preliminary studies, our team will develop innovative statistical methods that enable the search for translated uORFs over the full transcriptome for the most comprehensive uORFs identification by jointly modelling the fine-scale structure in ribosome profiling data around translated uORFs and coding regions.

Research plan

Aim 1: Develop novel statistical methods that leverage the fine-scale structure in ribosome profiling data to comprehensively identify translated uORFs using human lymphoblastoid cell lines as an initial model system.

Previous analyses [1, 3-5] have identified translated uORFs under certain restrictions, for example, 1) assuming that the translation of uORFs start at only canonical start codons or 2) restricting analysis to upstream of translated coding regions. Here, we propose to build on our recent work [1] to develop novel methods that will effectively address those issues from the previous analyses and enable comprehensive identification of translated uORFs.

Aim 1a. Jointly model the fine-scale structure in ribosome profiling data around translated uORFs and coding region, and learn the structure from the data.

First, we propose to build a model that 1) captures two distinct features of ribosome profiling data within translated regions, i.e., higher abundance and three-base periodicity (Figure 2A), but 2) allows translated uORFs and coding regions to have different fine-scale structures. For example, codons in uORFs and coding regions can have different three-base periodicity patterns in the data, or uORFs and coding regions can have different translation rates, i.e., ribosome profiling counts within translated coding regions are higher/lower than those within translated uORFs. We propose to learn the fine-scale structures from the data in order to better identify translated uORFs and coding regions, which also provides flexibility to capture potential differences in the structure between uORFs and coding regions.

Hidden Markov model: Specifically, we will model the data for each transcript using a hidden Markov model (HMM), as illustrated in Figure 3. Each base belongs to one of the 21 hidden states – 5'U (5' untranslated state), uI1, uI2, uI3 (first, second, third bases in a codon for translation initiation at uORF), uE1, uE2, uE3 (first, second, third bases in a codon for translation elongation at uORF), uT1, uT2, uT3 (first, second, third bases in a codon for translation termination at uORF), 5'U2 (untranslated bases downstream of uORF), I1, I2, I3 (first, second, third bases in a codon for translation initiation at coding region), E1, E2, E3 (first, second, third bases in a codon for translation elongation at coding region), T1, T2, T3 (first, second, third bases in a codon for translation termination at coding region), 3'U (3' untranslated state). The states {uI1, uI2, uI3, uE1, uE2, uE3, uT1, uT2, uT3} denote translated bases in a uORF and the states {I1, I2, I3, E1, E2, E3, T1, T2, T3} denote translated bases in a coding region. The remaining states denote untranslated bases. The groups of states {uI1, uI2, uI3}, {uT1, uT2, uT3}, {I1, I2, I3}, and {T1, T2, T3} help capture substantially higher ribosome profiling counts within codons at translation initiation/termination sites (Figure 2A).

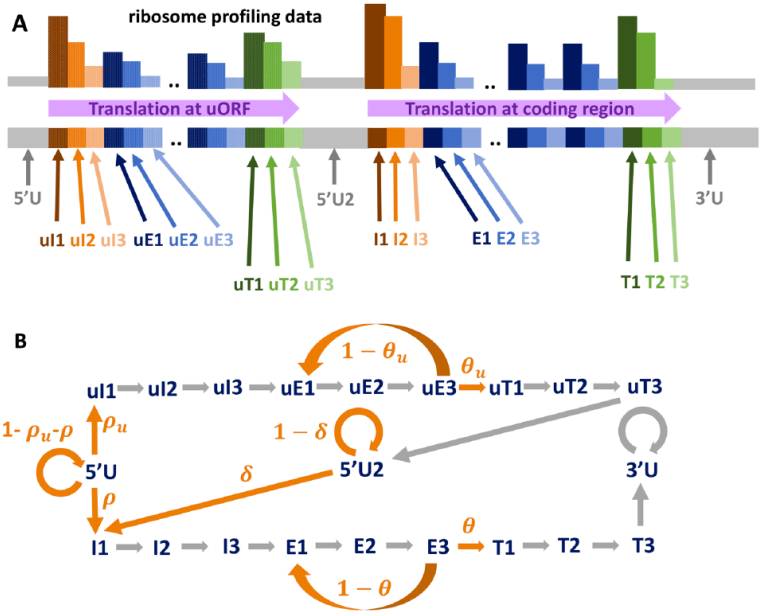

Figure 3 Model illustration *(A) Each base within a transcript belongs to one of the 21 hidden states. (B) Transitions with nonzero probabilities are indicated by arrows, with grey arrows denoting a probability of 1 and orange arrows denoting probabilities that are a function of the underlying sequence.*

Transition probability in a hidden Markov model: To allow each transcript to have 1) both a translated uORF and a translated coding region, 2) only a translated uORF, 3) only a translated coding region, or 4) none of them, we will consider possible transitions between hidden states as shown in Figure 3B. For example, if along a transcript, a hidden state starts from 5' untranslated region (5'U), moves to a translation initiation site for coding region (I1) with probability ρ, moves through translated bases in coding region (I2, I3, E1, E2, E3, T1, T2, T3), and finally ends at 3' untranslated region (3'U), the transcript has only a translated coding region.

We will allow translation for uORFs and coding regions to start at non-canonical start codons by modelling the transition probabilities ρ_u (corresponding to start of translation for uORF), and ρ and δ (corresponding to start of translation for

coding region) as a function of underlying RNA sequence at a given position. In addition, we can easily incorporate additional information, such as RNA sequence context [17] and RNA structure [18], that possibly affects translation initiation or termination, by modelling the transition probabilities as a function of additional information.

Emission probability in a hidden Markov model: Conditional on the state assignment, we will model ribosome profiling count at a given base, to account for three-base periodicity in the counts within translated regions, and the observation that translated base positions have a higher average count compared to untranslated base positions. Moreover, we will explicitly account for differences in the ribosome profiling count due to differences in transcript expression levels by using transcript-level RNA-seq data (or ribosome profiling data if RNA-seq data is not available) as a normalization factor.

Identification of translated uORFs using a hidden Markov model: We will learn the fine scale-structure from data by estimating the model parameters from the data over multiple transcripts. We will use an Expectation–Maximization algorithm to compute the maximum likelihood estimates for the model parameters. Using these parameters, we will infer the maximum a posteriori (MAP) hidden state sequence for each of the transcripts. We will retain transcripts whose MAP state sequence contains a pair of translation initiation and termination sites for uORF and has a high posterior probability. We will identify translated uORFs using the MAP state sequence.

Potential extension of the model: The proposed model has potential for further extension. For example, we can consider a nonzero transition probability from 5'U2 to uI1 (or from 3'U to I1) in order to allow each transcript to have more than one translated uORF (or coding region). Also, when prior knowledge of the structure is available (e.g., length of translated uORF is typically shorter than that of translated coding region), we can incorporate it into the model by putting constraints on the model parameters.

Aim 1b. Annotate a complete set of translated uORFs in human lymphoblastoid cell lines.
We will apply our proposed methods to identify translated uORFs in human lymphoblastoid cell lines (LCLs) for which gene expression phenotypes were measured genome-wide: mRNA in 86 individuals, ribosome occupancy in 72 individuals and protein levels in 60 individuals [19, 20]. In our previous study [1], we already assembled over 2.8 billion RNA sequencing reads into transcripts. This assembly gives us annotated transcripts that are expressed in LCLs, and we will restrict our analysis to transcripts with at least five ribosome profiling reads mapped to each exon. We will estimate the maximum likelihood estimates of the model parameters using highly expressed genes. Using these parameters, we will infer the MAP hidden state sequence for each of the transcripts, and identify translated uORFs as described in Aim 1a.

We will assess our proposed methods by comparing the identified translated uORFs with the ones identified in our previous study [1]. We expect the proposed methods to identify most of those from the previous study and additional new translated uORFs, such as those that have no translation in downstream coding regions. We will further assess our methods by experimentally validating the identified translated uORFs in Aim 2.

Aim 1c. Implement the proposed methods in freely-available open-source software for end-users to identify translated uORFs in their tissues/organisms of interests, and build a web-based platform for effective sharing and visualization.
Methods developed in Aim 1a will be implemented in user-friendly software and distributed freely to the academic community. We will use the statistical package R for initial development, and anticipate releasing many of the methods as R packages, with parts of the code that need optimization being performed in C/C++, and scripts for data processing in Perl/Python. If it becomes desirable to release some methods as a stand-alone C/C++ package, then our team has ample experience with this. Code will be thoroughly tested in house. CI Shim and PI Pique-Regi have an extensive history of successful software development and sharing. CI Shim is the developer and maintainer of six software packages (multiseq, SUCcESS, TileHGMM as R packages, WaveQTL, mvBIMBAM, BayesCAT as stand-alone C/C++ packages) and PI Pique-Regi is the developer and maintainer of three software packages (QuASAR, CENTIPEDE as R packages, GADA as stand-alone C/C++ packages).

Moreover, we will build a web-based platform to effectively share and visualize genome-scale results from Aim 1b. We will use the UCSC browser hub by loading our custom tracks into the Genome Browser. If it becomes desirable to have a browser embedded in our website, we will build an embeddable interactive genome visualization by using igv.js - JavaScript implementation of the Integrative Genomics Viewer (IGV). PI Pique-Regi has ample experience with building a web-based platform (CentiSNPs and GxE browser). We will also use Shiny in RStudio in order to display interactive plots and tables.

Aim 2: Experimentally validate translated uORFs using proteomics and next generation sequencing. The goal of this aim is twofold. First, we want to validate the translated uORF identified by the proposed method. Second, we want to use the validation rate to evaluate the method performance and to further identify the optimal parameters for the method. We will perform two independent lines of validation.

Aim 2a. Use tandem mass spectrometry to identify peptides produced from the translated uORFs. The major challenge in validating peptide production from translated regions is rooted in the size of translated regions. Based on our previous study, the average length of the peptides produced from translated uORFs are ~23 amino acids [1]. These peptides range in molecular weight between ~2 to 5 kilo Dalton (KD), which is much smaller than the detection range of most typical mass spectrometry experiments.

Preliminary experiment: In order to isolate and enrich for these smaller peptides from the cell lysate, we adapted the C8 column protocol published by Ma et al. [21]. To do so, we first replicated the mass spectrometry study on small open reading frame encoded peptides published by Ma et al. to bench mark our protocol. Protein extract prepared from K562 cells was processed by solid phase extraction using C8 extraction cartridge. This method successfully enriched for proteins with smaller molecular weight (Figure 4). We then performed LC-MS/MS for protein identification. From the C8 extracted protein samples, our mass spectrometry experiment identified 2,240 proteins with a median size of 35 KD (Figure 4C). Among the proteins identified using this method 85 (i.e. 4%) were smaller than 10 KD. With this purification protocol, we are now better equipped to validate peptide production from uORFs. Currently, we are further optimizing the protocol with applications of molecular weight filters and additional gel isolation steps to home in on the lower molecular weight targets.

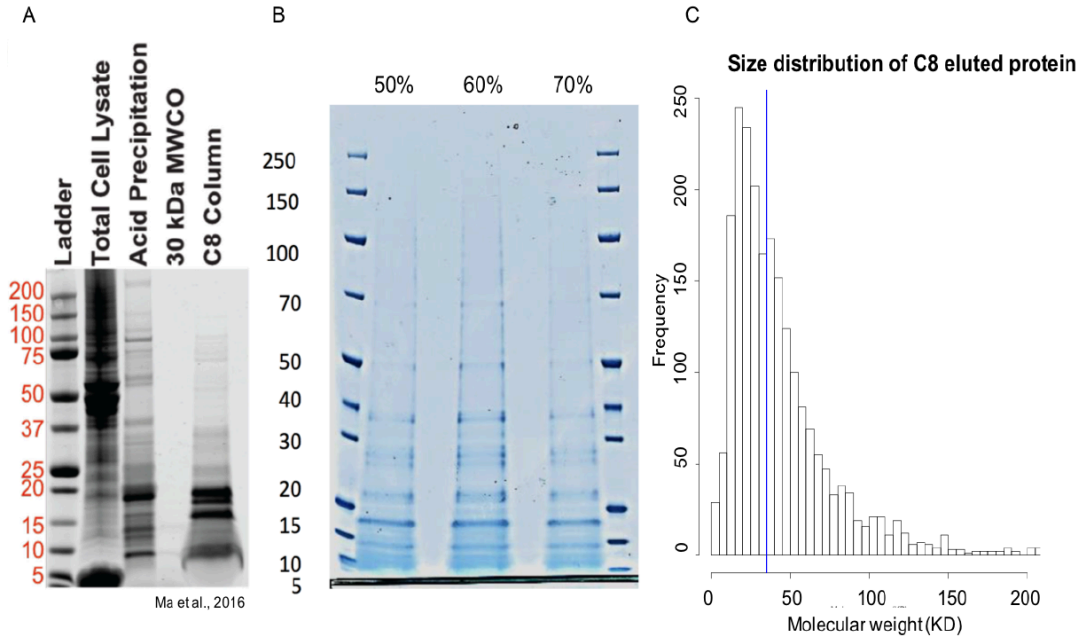

Figure 4 *C8 cartridge solid phase extraction successfully enriched for proteins of smaller molecular weights. (A) Results from Ma et al. comparing size distribution of proteins between whole cell lysate and C8 extracted protein samples. (B) Our SDS-PAGE results from replicating the Ma et al experiment with different elution conditions. Percentage of acetonitrile buffer used for elution are labeled at the top. Size distributions similar to Ma et al. results were seen in all three elution conditions (size ladders are labeled in KD). (C) Size distribution histogram of C8 extracted proteins that were successfully detected in LC-MS/MS experiments. The blue vertical line indicates the median molecular weight of all detected proteins. The x-axis is truncated in order to make the majority of the data points legible.*

Protein identification: We plan to perform protein identification from cell lysates prepared using our optimized protocol. We will perform six biological replications, each composed of cell lysates prepared from 12 non-

overlapping Yoruba individuals to cover the total 72 individual Yoruba cell lines used for our analysis in Aim 1b. We will perform peptide sequence search on the mass spectrometry data using Andromeda [22] supplemented with a custom peptide sequence database generated from the translated uORF identified in Aim 1b. We will calculate a peptide matching rate for translated uORF as a metric for method evaluation. To provide a better interpretability to the matching rate, we will compare the rate to the matching rate of currently annotated proteins, with appropriate adjustment of expression level and sequence composition for a fair comparison [1].

Aim 2b. Use harringtonine-treated ribosome profiling data to validate translation initiation at the identified uORF.
We will use an independent set of ribosome profiling data to validate the translated uORF identified in Aim 1b. Although peptide production is the ultimate standard for validating translation events, peptides with short half-life and certain amino acid compositions are much harder to validate using mass spectrometry. A validation rate obtained using mass spectrometry data is therefore often an underestimate. We have previously performed independent ribosome profiling experiments on a selected subset of cell lines with a harringtonine treatment step in the protocol [1]. Harringtonine treatment arrests the translation initiation complex and leads to a signature enrichment peak at the translation initiation site in the corresponding ribosome profiling data. We will use the presence (and absence) of such a signature peak at the translated uORFs identified by our method to calculate a validation rate. We will adapt peak calling algorithms from ChIP-seq and ChIP-exo data analysis to identify a translation initiation peak at high-resolution. We will further account for local mappability issues of the genome and potential sparsity issues in the data to adjust for background level. As a validation rate, we will report the proportion of the identified translated uORFs having a peak at harringtonine-treated ribosome profiling data.

FEASIBILITY

	2020		2021		2022	
Aim 1a: Develop methods	■	■				
Aim 1b: Annotate translated uORFs on human LCLs		■	■			
Aim 1c: Software development			■	■		
Aim 1c: Build a web-based platform					■	■
Aim 2a: Validate using tandem mass spectrometry				■	■	
Aim 2b: Validate using harringtonine-treated data				■		
All aims : Project dissemination				■	■	■

Table 1: The proposed timeline for this project

This project has been carefully developed to ensure the research can be successfully completed within the proposed timeline (Table 1) and requested budget. The project is highly-achievable given the team's demonstrated expertise, and collaborative and supportive research environment. This project will apply our team's expertise in statistical modelling of fine-scale structures in high-throughput sequencing data (Shim), gene regulation at the levels of transcription and translation (Shim, Wang, Pique-Regi), computational genomics (Pique-Regi), proteomics (Wang), and software development (Shim and Pique-Regi) for identification of translated uORFs. Particularly, the approach to the project builds on our team's successful work published in eLife where we identified and experimentally validated translated coding regions by modelling the fine-scale structure in ribosome profiling data. The employment of one research associate to pursue the project goals under the close guidance of our team will ensure the aims are fully addressed in a timely and efficient manner. PI Pique-Regi's other grant responsibilities are primarily of an oversight and training nature and compatible with his commitment to this proposal. The data are publicly available. The ribosome profiling data from human LCLs was collected by PI Wang [1] and previously analysed by our team in multiple projects [1, 9]. Validating peptide production from short translated uORFs using typical mass spectrometry experiments is challenging. However, we already have preliminary experimental results in order to isolate and enrich for small peptides from the cell lysate, and we are currently further optimizing the

protocol (Aim 2a). PI Wang will design the proposed mass spectrometry experiment, and we will outsource it to MSBioworks with which PI Wang has established a long-term working relationship. The research associate will analyse the mass-spec data for validation under the guidance of CI Shim and PI Wang. The research associate will visit PI Wang's group at the University of Texas Health Science Center at Houston for two weeks in July, 2020 to discuss and learn analysis for peptide sequence search using custom database.

The research associate and the PhD student will be based in CI Shim's group at the Melbourne Integrative Genomics (MIG) and the School of Mathematics and Statistics. Thus, trainees will be surrounded by other researchers working at the interface of statistics, bioinformatics, and computational biology, and will have excellent opportunity to get informal guidance with the method development and data analysis. Also, the MIG provides CI Shim's group with high-performance computing cluster and computational support such as a genomic data specialist.

BENEFIT

The understanding gene regulatory mechanisms is the cornerstone for translational studies. This project will provide the most comprehensive annotation of translated uORFs in human LCLs, and the methods for end-users to identify translated uORFs in their tissues/organisms of interests. Those outcomes will provide biological insight into uORFs as translational regulatory elements, facilitating progress in understanding translational regulatory mechanism. Therefore, the proposed project will add further knowledge to the research field of gene regulation. This project is highly beneficial for Australia to maintain its foothold in basic molecular biological sciences as well as position it at the forefront of this rapidly moving field of functional genomics.

This project is highly cost-effective. Publicly available ribosome profiling data sets have been mainly used to identify translated coding regions. The proposed methods will enable scientists to exploit those existing data sets to tackle new aim, identification of translated uORFs.

Another important benefit will be the training of early career scientists (the research associate and PhD student) in this rapidly moving field of functional genomics. They will receive training in the interface of statistics, computational biology, and bioinformatics, which will position them as future leaders of multi-disciplinary research fields.

COMMUNICATION OF RESULTS

During and upon completion of the proposed project, the results will be communicated to the broad community of researchers through 1) publications in leading journals (e.g. eLife, Nature Methods, PLoS Computational Biology), 2) presentations at appropriate national and international conferences (e.g., GeneMappers 2020 in Australia, Probabilistic Modelling in Genomics 2021 in Europe, Australasian Genomic Technologies Association conference 2021 in Australia, Biology of Genomes 2022 in US, International Conference on Research in Computational Biology 2022), and 3) a publicly available browser that we will build for effective sharing, searching, and visualization of results. The software implementing the proposed methods will be released as freely-available open-source software and maintained by the team. We will continue to communicate our work to the broader community through press releases and local media.

MANAGEMENT OF DATA

This project will produce annotation of translated uORFs and coding regions and mass spec validation dataset on human LCLs. Those data will be stored on a MIG (Melbourne Integrative Genomics) owned, Melbourne Bioinformatics administered high performance computing cluster with adherence to a data management policy. University of Melbourne has a data management policy in place (https://policy.unimelb.edu.au/MPF1242 "Management of Research Data and Record Policy") specifying all data will be managed per legal, statutory, ethical and ARC requirements. Following publication, these data will be made freely available to the broad research community in a browser which we will build as a part of the project. The proposed methods will be implemented as freely-available open-source software. The software and scripts implementing our methods and analyses will be made freely available in a public repository, such as GitHub (as is current practice for our team).

REFERENCES

1. Raj, A et al. 2016. Thousands of novel translated open reading frames in humans inferred by ribosome footprint profiling. eLife 5:e13328
2. Ingolia, NT et al. 2009. Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling.

Science 324:218–223.
3. Fields, AP et al. 2015. A Regression-Based Analysis of Ribosome-Profiling Data Reveals a Conserved Complexity to Mammalian Translation. Mollecular Cell 60, Issue 5, 816–827
4. Ji, Zhe et al. 2015. Many lncRNAs, 5'UTRs, and pseudogenes are translated and some are likely to express functional proteins. eLife. e.08890
5. Calviello, L et al. 2016. Detecting actively translated open reading frames in ribosome profiling data. Nature Methods 13, 165–170
6. Barbosa, C et al. 2013. Gene Expression Regulation by Upstream Open Reading Frames and Human Disease. PLoS Genet 9(8): e1003529
7. Morris, DR et al. 2000. Upstream Open Reading Frames as Regulators of mRNA Translation. Mol. Cell. Biol. vol. 20 no. 23 8635-8642
8. Calvo, SE et al. 2009. Upstream open reading frames cause widespread reduction of protein expression and are polymorphic among humans. PNAS. 106 (18) 7507-7512.
9. Battle, A et al. 2015. Impact of regulatory variation from RNA to protein. Science 347 (6222), 664-667
10. Li, JJ et al. 2015. Gene expression. Statistics requantitates the central dogma. Science. 6;347(6226):1066-7
11. Laurent JM et al. 2010. Protein abundances are more conserved than mRNA abundances across diverse taxa. Proteomics. 10(23):4209-12
12. Chick, JM et al. 2016. Defining the consequences of genetic variation on a proteome-wide scale. Nature 534: 500–505
13. Arcondéguy, T et al. 2013. VEGF-A mRNA processing, stability and translation: a paradigm for intricate regulation of gene expression at the post-transcriptional level. Nucleic Acids Res. 41(17):7997-8010.
14. Starck SR et al. 2016. Translation from the 5′ untranslated region shapes the integrated stress response. Science. 351(6272): aad3867.
15. Ingolia NT et al. 2011. Ribosome profiling of mouse embryonic stem cells reveals the complexity and dynamics of mammalian proteomes. Cell 147:789–802.
16. Lee S et al. 2012. Global mapping of translation initiation sites in mammalian cells at single-nucleotide resolution. Proceedings of the National Academy of Sciences of the United States of America. 109:E2424–E2432.
17. Kozak M. 1987. An analysis of 5'-noncoding sequences from 699 vertebrate messenger RNAs. Nucleic Acids Research 15:8125–8148.
18. Ouyang Z et al. 2013. SeqFold: genome-scale reconstruction of RNA secondary structure integrating high-throughput sequencing data. Genome Research. 23(2):377-87.
19. Pickrell JK et al. 2010. Understanding mechanisms underlying human gene expression variation with RNA sequencing. Nature. 464:768-72.
20. Battle A et al. 2015. Genomic variation. Impact of regulatory variation from RNA to protein. Science 347:664–667.
21. Ma J et al. 2016. Improved Identification and Analysis of Small Open Reading Frame Encoded Polypeptides. *Anal Chem*. 88: 3967–3975.
22. Cox J et al. 2011. Andromeda: a peptide search engine integrated into the MaxQuant environment. J Proteome Res. 10(4):1794-805.