UNIVERSITY OF TARTU

Faculty of Science and Technology

Institute of Bioengineering

Sviatoslav-Oleh Savchak

Identifying microproteins with genetic association data

Bachelor's Thesis (12 ECTS)

Curriculum Science & Technology

Supervisor(s):

Junior Research Fellow, MSc Anastasiia Alekseienko

Research Fellow in Functional Genomics, PhD Erik Abner

Tartu 2025

Identifying microproteins with genetic association data

Abstract

The annotation of the human genome is an ongoing process, continually refined as new functional elements are discovered. In recent years, regions previously classified as long non-coding RNAs have gained attention as many of them contain sequences that are actively translated. In this work, we mapped single-nucleotide variants (SNVs) listed in the GWAS Catalog to 7,264 previously undescribed open reading frames (ORFs) identified from the human genome. From this analysis, we identified six final microprotein encoding candidates that are likely to be associated with diseases or traits linked to their corresponding SNVs. These findings will hopefully contribute to the expanding functional annotation of the human genome and highlight the potential biological relevance of previously overlooked potential new microgenes.

Keywords: Genome-wide association studies, microproteins

CERCS: B110 Bioinformatics, medical informatics, biomathematics, biometrics

Institute name: Institute of Genomics, University of Tartu

Research group: Functional Genomics

Mikrovalkude tuvastamine geneetiliste assotsiatsiooniuuringute abil

Lühikokkuvõte

Inimgenoomi annotatsioon on pidevalt arenev protsess, mida täiustatakse uute funktsionaalsete elementide avastamisel. Viimastel aastatel on kasvanud huvi nende genoomipiirkondade vastu, mida varem peeti pikkadeks mittekodeerivateks RNA-deks, kuid mis sisaldavad aktiivselt transleeritavaid järjestusi. Käesolevas töös kaardistasime inimese genoomist tuvastatud GWAS catalog andmebaasis loetletud üksiknukleotiidseid polümorfisme (SNV-sid) 7 264 varem kirjeldamata avatud lugemisraamide sees. Selle analüüsi tulemusena tuvastasime kuus mikrovalku kodeerivat alad, mis on tõenäoliselt seotud haiguste või tunnustega, mis seonduvad vastavate SNV-dega. Need tulemused võivad aidata kaasa inimese genoomi funktsionaalse annotatsiooni täiendamisele ning toovad esile varasemalt tähelepanuta jäänud potentsiaalsete uute geenide bioloogilise tähtsuse.

Võtmesõnad: Geneetilised assotsiatsiooniuuringud, mikrovalgud

CERCS: B110 Bioinformaatika, meditsiiniinformaatika, biomatemaatika, biomeetrika

TABLE OF CONTENTS

TERMS, ABBREVIATIONS, AND NOTATIONS

GWAS	–	Genome-wide association study
lncRNA	–	long non-coding RNA
ORF	–	open reading frame
Ribo-seq	–	Ribosome sequencing/profiling
SNV	–	single-nucleotide variant
sORF	–	short open reading frame
CADD	–	Combined Annotation Dependent Depletion

INTRODUCTION

More than 20 years have passed since the completion of the first drafts of the human genome, and its annotation has been continuously updated. However, the exact number of human genes remains undecided. One group of proteins has been especially overlooked throughout these years: microproteins, which are polypeptides encoded by short open reading frames (sORFs) less than 300 nucleotides long. Advances in ribosome profiling (Ribo-seq) have shown that sORFs that were previously characterized as non-coding sequences are actually being translated in human cells. Recently, a major reference annotation project was created to standardize a set of microproteins to facilitate research for discovering their functions in the global community.

Considering this knowledge gap in terms of human genome annotation, this work aims to determine microproteins' involvement in human common diseases or phenotypic traits. To discover such connections, a set of statistically significant single-nucleotide variants (SNVs) was acquired from the GWAS Catalog. The core idea of a genome-wide association study (GWAS) is to discover genetic variants that are statistically correlated with a studied disease or trait.

By focusing on validated phenotype-associated SNVs located within sORFs, we aim to highlight candidates with potential functional relevance. These SNVs are further prioritized using CADD (Combined Annotation Dependent Depletion) scores, which estimate the deleteriousness of the genetic variants. Our approach provides a template for hypothesizing the function of uncharacterized microproteins based on their potential involvement in human phenotype or disease susceptibility, contributing to a more complete annotation of the human genome.

1 LITERATURE REVIEW

1.1 Open reading frames

Converting raw genomic sequences into biologically meaningful information in the context of the central dogma of molecular biology is the primary goal of genome annotation. It involves taking genomic data - DNA or RNA sequences - and identifying functional elements (genes) in the correct genome locations.

1.1.1 Open reading frames in genome annotation

A key component of genome annotation is the identification of open reading frames (ORFs) - stretches of nucleotide sequences with the potential to be translated into proteins. To understand what constitutes an ORF, it is important to consider the reading frame – one of six (three per strand) possible ways to divide a given double-stranded nucleotide sequence into codons (triplets of nucleotides coding for a specific amino acid or stop signal). Within a given reading frame, an ORF refers to a region that is not interrupted by a stop codon (Sieber *et al,* 2018). However, it can be bounded differently according to a particular definition (Figure 1). The most common interpretation of ORFs aligns them with a coding DNA sequence (CDS), which is a region flanked by start and stop codons and known to be translated into a functional protein (Figure 1 - Definition 1).

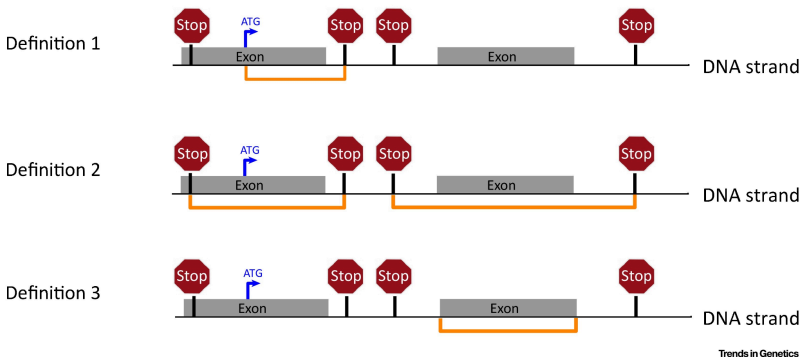

Figure 1. Three different definitions yield different ORFs (highlighted by orange lines) in eukaryotes. An ORF is bounded by: start and stop codons (Definition 1); two stop codons (Definition 2); donor and acceptor splice sites (Definition 3) (Adapted from Sieber *et al,* 2018)

Since the early human genome annotations, identifying functional short ORFs (sORFs) and their putative protein products has been a challenging task for researchers; therefore, the current annotation projects are missing most of them (Basrai *et al*, 1997; Saghatelian & Couso, 2015). The difficulty is caused by numerous start and stop codons randomly occurring in the genome and producing non-coding sequences, which serve as a high noise that prevents functional, protein-coding sORFs from being readily recognized (Basrai *et al*, 1997; Saghatelian & Couso, 2015). Hence, particular criteria were applied to define ORFs for the simplification of the computations, such as 1) conventional start (i.e., AUG) and stop codon (i.e., UAA, UGA, and UGA), 2) single ORF per transcript, 3) length of at least 100 codons (300 nucleotides) (Schlesinger & Elsässer, 2022). As a result of the third criterion, the abundance of annotated human proteins declines rapidly below 100 aa (Figure 2) (Frith *et al*, 2006).

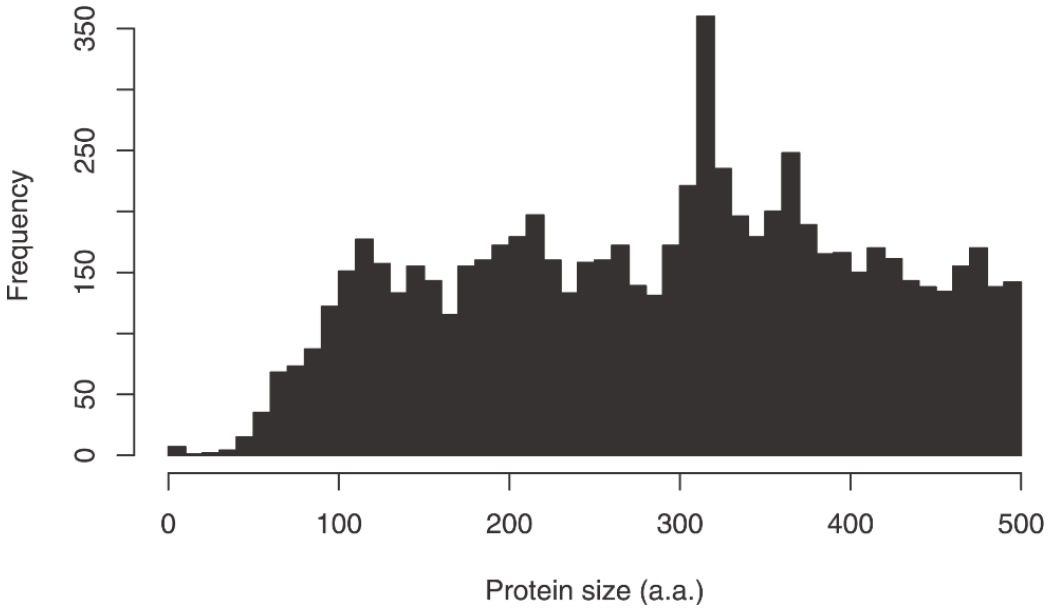

Figure 2. Size distribution of 11,679 human proteins from Swiss-Prot. Only 405 proteins (3.5%) are shorter than 100 aa. (Adapted from Frith *et al*, 2006)

1.1.2 Reasons behind ORF length cut-off

The ORFs that are smaller than a 300-nucleotide (100-codon) threshold and their corresponding protein product are referred to as sORFs and microproteins (but also known as micropeptides or sORF-encoded protein (SEP)), respectively. Although many known microproteins possess essential functions, there are particular reasons behind setting a 300-nucleotide (100-codon) cut-off (Leong *et al*, 2022).

Firstly, the likelihood of a start codon encounter in a nucleotide space is 1 in 64, while the probability of a stop codon generation within the following 99 codons is approximately 99%. This implies that approximately 1.5% of the human genome could have encoded for ORFs < 100 codons. (Olexiouk *et al*, 2018). The chances of this excessive amount of putative protein-coding ORFs being translated into functional proteins appear remote (Leong *et al*, 2022). Another reason for introducing the cut-off is that conventional algorithms are unsuitable for annotating the sORFs. For instance, evolutionary conservation is a strong indicator of functionality, and short sequences tend to be granted low scores in the analysis (the shorter the region of similarity, the greater the chance it could have occurred randomly. (Ladoukakis *et al*, 2011). Finally, it is associated with technical difficulties in acquiring experimental proof for the sORF function. Standard biochemical techniques for protein isolation cannot readily detect small peptides (below 10 kDa), which can elude a typical gel or filter and be obscured by larger protein degradation products. Genetic approaches also face difficulties since the short length of sORF makes them small targets for random mutagenesis and other gene-discovery protocols. The sheer amount of sORFs in the genome makes systematic directed mutagenesis impractical. Even when a sORF mutation is isolated, it is often attributed to adjacent canonical genes, as most sORFs remain unannotated (Couso & Patraquim, 2017).

Due to the aforementioned reasons, past researchers typically excluded any ORFs that yield proteins smaller than 100 amino acids (aa) in eukaryotes. Hence, the number of annotated microproteins is much smaller than it is predicted to be (Frith *et al*, 2006).

1.2 Microproteins

The emergence of Riboseq profiling (Ribo-seq) led to a reevaluation and hence an alternative interpretation of long non-coding RNAs (lncRNAs), which may actually contain numerous functional ORFs within their sequences (Mudge *et al*, 2022). Combined with mass spectrometry (MS), these methods have narrowed down this large number of likely coding sORFs to various thousands (Schlesinger & Elsässer, 2022).

Consequently, scientists are starting to explore the functions of microproteins. However, the lack of evidence of their experimental reproducibility and physiological roles raised a controversy about featuring them as functional genes in reference annotation databases.

1.2.1 Ribosome Profiling

Ribo-seq is a deep sequencing-based tool that provides comprehensive and detailed translation measurements in vivo (Ingolia *et al*, 2009). This technique is performed by isolating ribosome-bound RNA fragments of ~30 nucleotide bases, offering evidence of translation of specific mRNA regions with single-nucleotide resolution (Makarewich & Olson, 2017).

The approach relies on the ability of translating ribosomes to protect RNA from nuclease-mediated degradation. Translation inhibitors are introduced into cytoplasmic lysates to halt active translation, and mRNA-ribosome complexes are treated with nucleases to produce ribosome-protected fragments (RPFs), commonly known as ribosome footprints (Figure 3). These footprints are then recovered and purified for further sequencing to identify the exact location of the ribosome at the time when translation was stopped. Combined with measuring RPF density on a transcript, the method enables gaining insights into the quantitative dynamics of translation and protein synthetic rates. Even though Ribo-seq efficiently detects translations, avoiding working with protein molecules that often bring difficulties, the ribosome occupancy does not prove the actual protein-coding potential and later functionality at the protein level (Mudge *et al*, 2022).

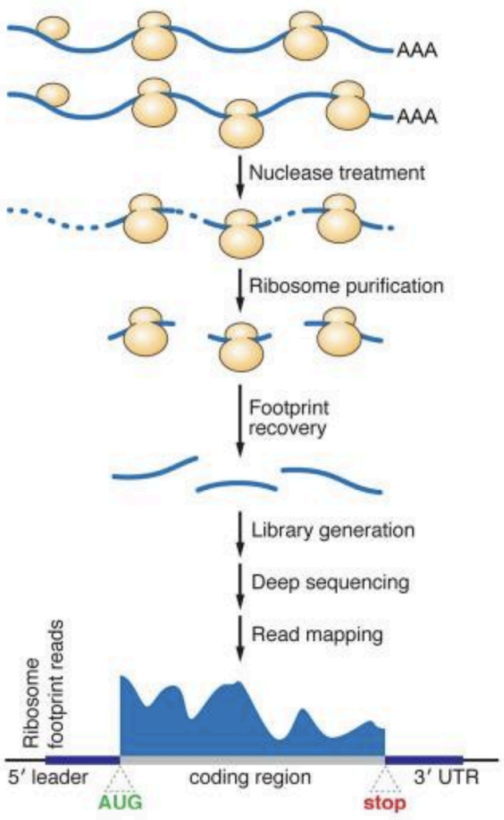

Figure 3. Ribosome profiling for the translation measurement. (Adapted from Makarewich & Olson, 2017).

Although Ribo-seq is an experimental method, assessing the coding potential of a region of interest is primarily a computational task. The obtained results expanded the existing understanding of the protein-coding potential of the genome, as they demonstrated that once considered non-coding RNA may actually comprise ORFs that are translated. In addition, canonical proteins showed an ability to be translated with alternative start and stop codons, along with the ORF capacity to be translated using non-conventional (non-AUG) start codons. (Ingolia, 2016)

1.2.2 Phase I dataset

In 2022, a group of reference gene annotation projects made a massive effort to create a standardized catalog of human Ribo-seq ORFs to facilitate global research and validation of the ORFs on the protein level (Mudge *et al*, 2022).

The authors selected seven ORF datasets from different human studies representing key projects for genome-wide Ribo-seq identification during the five years preceding the start of the research (Van Heesch *et al*, 2019; Ji *et al*, 2015; Calviello *et al*, 2016; Martinez *et al*, 2020; Chen *et al*, 2020; Gaertner *et al*, 2020; Raj *et al,* 2016). The selection of literature sources was based on the dataset's comprehensiveness, concentration on large-scale ORF detection, and transparency in reporting multiple categories of ORFs in the data files (Mudge *et al*, 2022). The total set of 39,788 translated ORFs corresponding to 29,373 unique protein sequences was collected from these publications.

The set was standardized to maximize the reproducibility across studies since each of them applied different minimum length cut-offs to define their Ribo-seq ORFs, and only four of the studies considered near-cognate ORFs (starting with a non-AUG start codon). Therefore, the ORFs smaller than 16 aa and those starting from non-AUG initiation codons were removed from the list. Furthermore, the researchers excluded all translated ORFs that could not be fully mapped to any transcript in Ensembl Release v.101 (August 2020, equivalent to GENCODE v35) transcriptome, resulting in 8,805 ORFs meeting such requirements (Figure 4). In addition, ORFs overlapping pseudogenes and in-frame complete coding sequences (CDS) were excluded. Finally, to get the non-redundant list of translated ORFs for Phase I, among ORF isoforms that share the same start and/or stop codon and at least 90% of the linear amino acid sequences, the longest ORF was selected as representative. The filtering resulted in a final Phase I consensus set of 7,264 Ribo-seq ORFs (Figure 4).

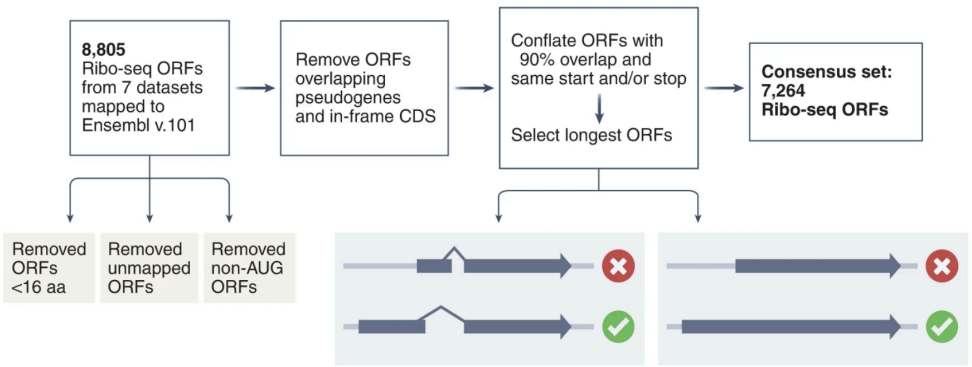

Figure 4. Schematic overview of filtering steps used to create the consensus set of Ribo-seq ORFs (Mudge *et al*, 2022)

Numerous ORFs tend to be found in more than one publication, totaling 3,085 such ORFs. However, despite demonstrating consistency in the Ribo-seq signal, it neither proves their biological function nor undermines the rest of the 4,179 non-replicated ORFs. The group published the dataset comprising these results and named the current state of the project Phase I.

In 2024, the same researchers published a preprint with Phase II of the Ribo-seq ORFs catalog to initiate a continuous endeavor to develop a consensus understanding of the protein-level evidence for 7,264 Ribo-seq ORFs. Following a search of 3.8 billion MS/MS spectra collected from immunopeptidomics datasets and regular protease digests from human cell lines, tissues, and fluids, the study confirmed that 1,715 Ribo-seq ORFs had strong proteomic evidence of translation, indicating that they should be investigated further (Deutsch *et al*, 2024)

1.3 SNVs and protein modifications

The human genome possesses a broad spectrum of genetic variation, from single-base pair changes to multimegabase cytogenetic alterations. (Sharp *et al*, 2006). Structural variations (SVs) are differences in large DNA segments across different genomes and are usually more than 100 base pairs (bp) long (Hollox *et al*, 2022). There are several types of SVs, such as duplication, deletion, and more complex tandem repeats, insertions, and inversions. There are a number of ways in which SVs can affect gene expression. The most straightforward is that duplication or deletion of entire genes can increase or decrease gene expression, respectively. Alternatively, SV might influence a gene's expression levels by changing the spatial arrangement between the gene and a regulatory element, as observed at the HoxD locus in mice (Montavon *et al*, 2012). SVs can produce novel fusion genes with new functions, such as immunity-related butyrophilins in humans, caused by a 56-kilobase deletion (Aigner *et al*, 2013). Despite the important role SVs play in contributing to phenotypic diversity and evolutionary adaptations, most studies in evolutionary genomics continue to focus more on single-nucleotide variants (SNVs) (Hollox *et al*, 2022).

1.3.1 Single-nucleotide variation

Single-nucleotide variations (SNVs) are one of the most frequent types of variation in DNA, accounting for about 90% of human genome variation (Collins *et al*, 1998). An SNV is a variation in a single nucleotide at a specific position within a genome. Human genomes differ from one individual to another due to these genetic variations, and therefore, they have become the primary focus of numerous disease-gene association studies. Sequencing of whole human genomes has revealed that the number of SNVs in each individual is approximately 3 to 5 million (Wheeler *et al, 2008*; Levy *et al,* 2007). Although most common SNVs are likely to be functionally neutral due to evolutionary forces (McClellan & King, 2010), some SNVs may influence biological functions and therefore play a direct role in disease susceptibilities and drug sensitivities. Identifying these functional SNVs represents a key objective of contemporary genetics and genomics studies.

Although SNVs may influence biological functions in various ways depending on their position in the genome, the primary focus of bioinformatics has been drawn to SNVs in protein-coding sequences. Those SNVs can be further divided into two classes: ones that alter the amino acid sequence and those that do not: non-synonymous (missense) and synonymous, respectively (Kaushik *et al*, 2019).

The effect caused by a missense SNV may be predicted based on two approaches. First, various metrics rely on the estimated evolutionary distance between the original and substitute amino acids (e.g., BLOSUM (Henikoff & Henikoff, 1992)). This approach considers amino acid substitution rates at the equivalent protein position in other species to assign a score to each mutation. Substitutions that follow evolutionary trends are less likely to be deleterious than substitutions that are less consistent with evolution. Another method is based on the physicochemical properties difference between each amino acid pair. It has been demonstrated that changes in protein hydrophobicity, net charge, packing density, and solvent accessibility correlate with the functional impact of missense SNVs (Wang & Moult, 2001).

1.3.2 Genome-wide association studies - steps

Genome-wide association studies (GWAS) explore genetic variations across the genomes of many individuals to discover genotype-phenotype links. While GWAS can analyze various

forms of human genome variation, SNVs are the most frequently examined genetic variants. A typical report acquired by GWAS is a set of SNVs that indicate statistically significant associations with a particular trait. A standard GWAS workflow consists of several steps (Uffelmann *et al*, 2021):

1. Data collection

Substantial sample sizes are required to obtain applicable genome-wide significant associations by leveraging GWAS. The data can be obtained from different sources depending on the required sample size, existing data availability, or the extent of efforts needed to acquire new data. Typical sources of data collection to conduct GWAS are biobanks or cohorts recruited based on the experimental goal, e.g., disease-focused.

2. Genotyping

Microarrays are the most common method for genotypic data acquisition microarrays due to their relative affordability. Although next-generation sequencing, such as whole-exome sequencing (WES) or whole-genome sequencing (WGS), is more expensive, it includes rare genetic variants in addition to common ones, contrary to microarrays.

3. Data processing

Careful quality control is required for obtaining reliable results. It includes measures at the wet-laboratory stage, such as genotype assay, and dry-laboratory stage, such as excluding low-confidence SNVs and principal component analysis (PCA).

4. Imputation

Omitted genotypes can be imputed using information from existing reference populations such as the 1000 Genomes Project or TOPMed (Taliun *et al*, 2021; Auton *et al*, 2015).

5. Association testing

Genetic association tests are conducted for each genetic variant to calculate its correlation with a specific trait. The appropriate statistical test can be selected based on the nature of the trait of interest (e.g., linear or logistic regression).

Genetic variants that are physically located close to each other tend to be in linkage disequilibrium (LD) and, therefore, not independent. To avoid false positive results, the genome-wide significance threshold of 5×10^{-8} was obtained by running numerous

simulations; it efficiently controls for the number of independent SNVs in the entire genome for studies on European populations (Marees *et al*, 2018).

6. Meta-analysis

Sometimes, it is recommended to perform association tests in unique cohorts separately and use appropriate methods to combine the results.

7. Replication

Results can be replicated using internal or external replication in an independent cohort. For external replication, the independent cohort must be ancestrally matched and have no shared individuals from the study cohort; this can indicate whether the results are statistically significant or obtained randomly in the study cohort.

8. Post-GWAS analyses

Various tools and resources are available for post-GWAS analysis to help interpret the association results from a functional or biological perspective and bring valuable insights.

The NHGRI-EBI GWAS Catalog is the largest and most comprehensive publicly accessible GWAS data, which contains hundreds of thousands of curated SNV-trait associations (Cerezo *et al*, 2025).

2 THE AIMS OF THE THESIS

The main aim of the thesis is to test the hypothesis that some unannotated micropeptides participate in common diseases or phenotypic traits.

We integrate genome-wide association data by selecting significant SNVs from the GWAS Catalog that are linked to common diseases or traits. We then assess whether these SNVs overlap with short open reading frames (sORFs), based on genomic coordinates, to identify potential connections between sORFs and human phenotypes.

3 EXPERIMENTAL PART

3.1 MATERIALS AND METHODS

3.1.1 Data sources

To address the study's objectives, two distinct datasets were required. The first dataset includes SNVs associated with phenotypes, whereas the second dataset consists of microprotein sequences and their corresponding genomic locations.

In this study, the GWAS Catalog v.1.0.2 (Cerezo *et al*, 2025) was used to obtain relevant genetic associations with phenotypic traits and common diseases. The dataset includes 786,898 SNVs gathered amongst 6,284 GWAS studies (as of 20th February 2025).

The Phase I consensus dataset, which consists of curated Ribo-seq ORFs, served as a primary source of microprotein data. It contains 7,264 human microproteins derived from seven independent Ribo-seq publications (Van Heesch *et al*, 2019; Ji *et al*, 2015; Calviello *et al*, 2016; Martinez *et al*, 2020; Chen *et al*, 2020; Gaertner *et al*, 2020; Raj *et al,* 2016).

3.1.2 Packages

The data handling and filtering were performed using Python v. 3.13.2 in a Conda virtual environment; pandas v. 2.2.3. The charts were done using numpy v. 2.2.3; matplotlib v. 3.10.0; and seaborn v. 0.13.2. The overlaying of two datasets was performed by using pyranges v. 0.1.4. Detecting changes in amino acid sequences was performed using pyfaidx v. 0.8.1.3 and biopython v. 1.85.

The Manhattan plot was generated using R v. 4.4.2; R topr v. 2.0.2; and R ggplot2 v. 3.5.2.

3.1.3 Data collection and cleanup

As the GWAS Catalog aggregates results from numerous independent studies, certain SNVs were duplicated, and the formatting varied across entries. To ensure consistency and retain a single copy of each unique SNV, the dataset was standardized – instances with invalid

chromosome IDs and chromosome positions were removed; the SNVs with the lowest p-value (i.e., the most significant) are retained, while other duplicates are excluded if present. In addition, retained SNVs were filtered by applying a genome-wide significance threshold of $5x10^{-8}$.

To increase the plausibility and robustness of the subsequent hypotheses, we retained only those microproteins that were reported by at least two independent Ribo-seq studies ("Riboseq_evidence" variable). In addition, the dataset included sequences that are equal to or longer than 300 nucleotides, which is not within the objectives of the work; hence, such instances were excluded from the dataset. Another layer of filtering is removing micropeptides that include introns (i.e., having multiple start and end positions).

3.1.4 Overlap

To identify SNVs located within microprotein-encoding genomic regions, two datasets were formatted to comply with PyRanges requirements by including the following columns:

- "Chromosome": chromosome identifier
- "Start": start coordinate of the genomic region
- "End": end coordinate of the genomic region

For the GWAS Catalog data, each SNV was represented as an interval by assigning its chromosomal position as both the start and end coordinates. The resulting subset of SNVs that overlap with the microprotein-encoding regions was then converted back into a pandas DataFrame for further analysis.

3.1.5 Variant prioritization

The variant prioritization was performed based on the estimated deleteriousness of each variant.

The corresponding rsIDs of each SNV were submitted to the Ensembl Variant Effect Predictor (VEP), which assigned a Combined Annotation Dependent Depletion (CADD) score to each variant. If multiple entries were returned for a single variant, the entry with the

highest “CADD_PHRED” value was retained to represent the most deleterious effect, while other duplicates were excluded. The obtained scores were then merged back with the original using the rsIDs.

Only variants with a CADD score above 12 were retained, representing the top 10% most deleterious variants in the human genome. This threshold was set to increase the likelihood of detecting meaningful associations between SNVs and potential microprotein function. In addition, VEP included a “Consequence” variable for each variant – SNVs classified as “missense_variant” (for “Consequence” variable) were excluded, as they indicate that these SNVs cause an amino acid change within an already annotated human gene. Since most microproteins are not listed as protein-coding genes, phenotypes associated with these variants are likely attributable to overlapping annotated genes rather than microproteins themselves.

3.2 RESULTS

The first stage of this thesis involved curating and refining two key datasets: genome-wide significant variants from the GWAS Catalog and predicted human microproteins from the Phase I sORF dataset. The GWAS Catalog dataset initially consisted of 786,898 statistically significantly associated SNVs. After standardization of the dataset, 632,656 SNVs remained. By applying the internationally agreed-upon genome-wide significance threshold of $5x10^{-8}$, the dataset was further reduced to 508,949 SNVs passing the cutoff.

Next, we cleaned and filtered the Phase I sORF dataset. The Phase I sORF dataset originally contains information about 7,264 predicted human microproteins. Retaining only those sequences that were found by at least two Ribo-seq studies resulted in only 3,085 microproteins fitting the criterion. The dataset initially included 131 proteins equal to or longer than 300 nucleotides, which were excluded as they did not meet the microprotein criterion of being 300 nucleotides or shorter. After filtering, 2,952 hypothetical microprotein-encoding regions remained. In addition, we retained only those regions that had a single genomic start and end coordinate (i.e., no introns), resulting in 1,997 such microproteins (Figure 5).

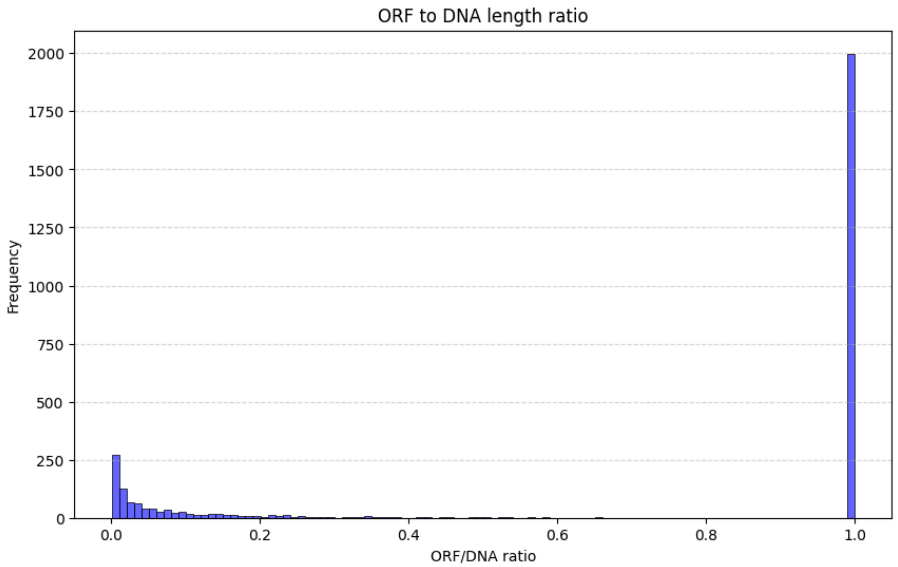

Figure 5. ORF to corresponding gene length ratios of 2,952 microproteins. Bin size is 0.01.

By merging the two datasets, we identified 2,838 SNVs located within the microprotein-encoding regions (Figure 6). To further explore the functional relevance of the variants, the processed and merged dataset was annotated. Subsequently, the unique rsIDs were extracted and submitted to the annotation tool VEP, allowing us to obtain CADD scores and map SNVs to known genes where applicable.

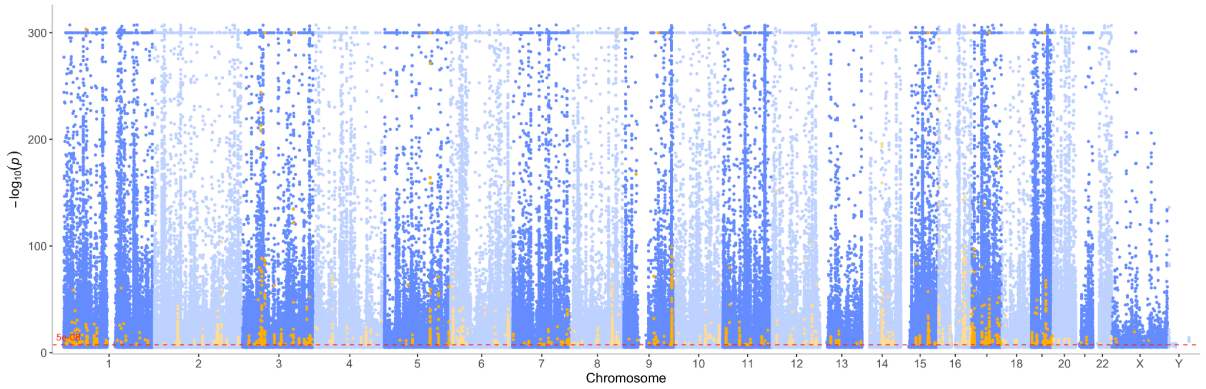

Figure 6. Manhattan plot of 632,656 standardized SNVs from the GWAS Catalog (blue) and 2,838 SNVs of interest (yellow). The horizontal dashed red line represents the genome-wide significance threshold at $p=5\times10^{-8}$ ($-\log_{10}(p) = \sim7.3$).

The variant prioritization resulted in 29 instances that were analyzed in-depth. Although it corresponded to only 14 unique SNVs, as the GWAS Catalog aggregates findings from multiple independent studies, a single SNV could have been reported multiple times. Only the most significant genetic association with a phenotype (the lowest p-value) was retained for each unique SNV.

To evaluate whether these variants alter the encoded microproteins, we assessed their potential to cause amino acid changes, as non-synonymous variants are more likely to result in functional or structural changes in the protein product. The alleles associated with each of the 14 unique SNVs were introduced into the human reference genome (Genome assembly GRCh38) for downstream analysis of the resulting ORF sequences (Figure 7). Of these, only seven SNVs were non-synonymous, leading to amino acid substitutions within encoded microproteins. Six of these substitutions occurred outside the initial amino acid groups, and these SNVs were selected as the final candidates.

Algorithm 1: Analyze SNP Impact on Protein Translation

Input: *genome*: Reference genome in FASTA format
microproteins: Table of microproteins and associated SNVs
Output: *results*: Dataset with microprotein variation details
Data: *wild_seq*, *mutated_seq*: DNA sequences
wild_protein, *mutated_protein*: Translated protein sequences

Function `process_snps`(*genome*, *remainings*)**:**
  Format CHR_ID in *microproteins* to match FASTA (e.g., "chr1")
    `// standardize chromosome names`
  Initialize empty list *results*
  foreach *row in microproteins* **do**
    end ← row[*ends*]
    start ← row[*starts*] − 1
      `// convert microprotein-encoding regions according to python index`
    snp_chr, *snp_pos* ← split row[$CHROM$] by "_"
      `// (e.g. 1_32817982)`
    risk_allele ← split row[$SNP - RISK\ ALLELE$] by "-" and take allele
      `// (e.g. rs3737251-C)`
    wild_seq ← genome[*snp_chr*][*start* : *end*]
      `// extract wild-type DNA sequence from` *genome*
      `// translate wild sequence`
    if *wild_seq*[0 : 3] == *"ATG"* **then**
      wild_protein ← Translate(*wild_seq*)
    else
      wild_protein ← Translate(ReverseComplement(*wild_seq*))
    rel_pos ← snp_pos − *start*
    Replace base at *rel_pos* in *wild_seq* with *risk_allele* → *mutated_seq*
    if *mutated_seq*[0 : 3] == *"ATG"* **then**
      mutated_protein ← Translate(*mutated_seq*)
    else
      mutated_protein ← Translate(ReverseComplement(*mutated_seq*))
      `// detect mutation in protein`
    if *wild_protein* ≠ *mutated_protein* **then**
      for *i* ← 0 **to** *length of wild_protein* **do**
        if *wild_protein*[*i*] ≠ *mutated_protein*[*i*] **then**
          mutation ← format "*i* + 1: *wild*[*i*] → *mutated*[*i*]"
          break
    else
      mutation ← "did not occur"
    Append mutation details to *results*
  return *results*

Figure 7. An algorithm for identifying amino acid changes in the microprotein sequences.

To explore the potential functional relevance of the final six candidates, we examined predicted structures of their wild-type sequences by using AlphaFold, as stable, well-structured proteins are more likely to have physiological roles or function as ligands in molecular interactions. However, AlphaFold struggles to produce confident structural predictions for some of the six candidates, likely due to their relatively small size (Figure 8). However, structure prediction for c18riboseqorf4 exhibited higher confidence; c17riboseqorf124 and c18riboseqorf4 revealed the presence of complex secondary structures, including alpha-helices.

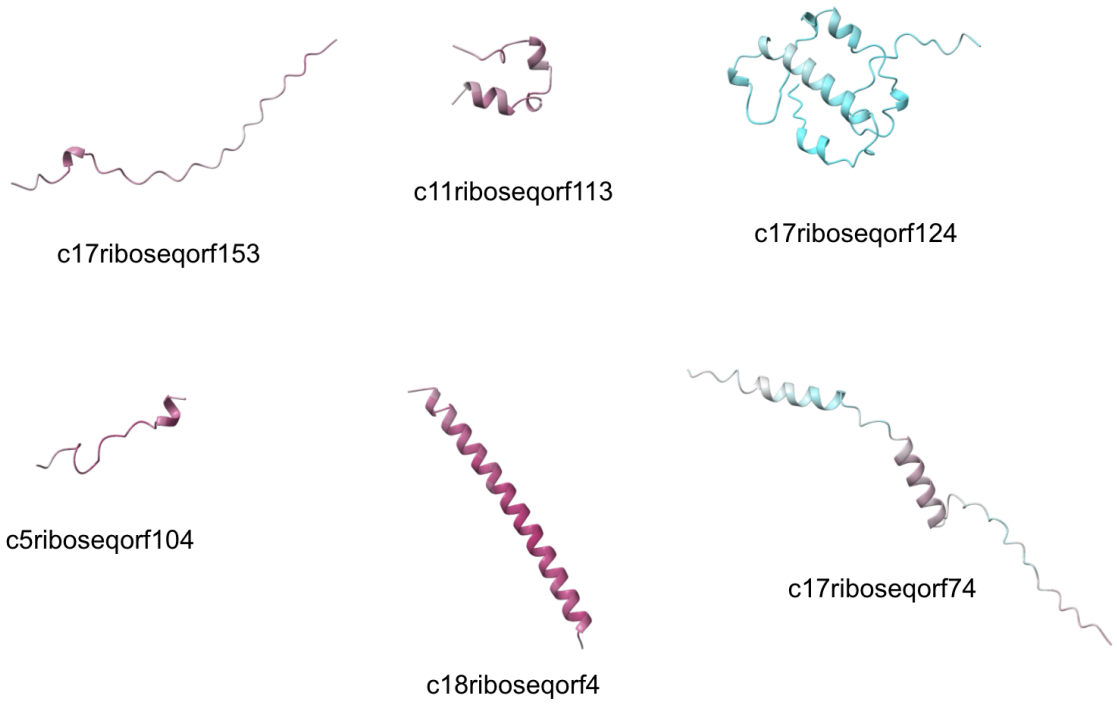

Figure 8. Predicted protein structures of the final 6 candidates, which were generated by AlphaFold. The color of the structure is determined by the pLDDT (predicted local distance difference test) value, indicating confidence in structure predictions (the value ranges from 0 to 100, from lowest to highest confidence). The palette used is "cyan-gray-maroon," ranging from lowest (cyan) to highest (maroon) confidence.

The generic summary of the six final candidates is included in Table 1. We manually analyzed genes that could be assigned to each Ribo-seq ORF and validated whether they potentially have a meaningful relationship to the associated trait/disease.

Table 1. Final 6 candidates. “Ribo-seq evidence” indicates how many Ribo-seq studies have identified the ORF. “Variation in protein sequence” indicates the changes in the amino acid sequences of the encoded microproteins (<amino acid position>:<original amino acid> → <altered amino acid>) and changes between amino groups of the original amino acid and the altered (<original group> → <altered group>).

Ribo-seq ORF name	ORF Length	Ribo-seq evidence	Variation in protein sequence	SNV rsID + Allele	p-value	Associated trait/disease
c5riboseqorf104	60	2	19:R → P Basic → Cyclic	rs1042711-C	2×10^{-23}	eosinophil (absolute count, mean, inv-norm transformed)
c11riboseqorf113	84	3	10:L → Q Aliphatic → Acidic/Amide	rs11820337-T	4×10^{-10}	Left-handedness
c17riboseqorf74	192	5	55:T → P Hydroxyl/Sulfur → Cyclic	rs17616365-G	3×10^{-23}	Sex hormone-binding globulin levels adjusted for BMI
c17riboseqorf124	291	2	59:P → R Cyclic → Basic	rs2070107-C	8×10^{-11}	Eosinophil percentage of white cells
c17riboseqorf153	99	2	3:P → T Cyclic → Hydroxyl/Sulfur	rs2247856-A	4×10^{-52}	Reticulocyte fraction of red cells
c18riboseqorf4	138	2	44:I → R Aliphatic → Basic	rs7811-G	6×10^{-13}	Blood pressure (pleiotropy model 2 SBP adjusted for estimated causal effects x DBP)

3.3 DISCUSSION

In this study, we present a novel approach that integrates genome-wide association study (GWAS) data with ribosome profiling (Ribo-seq)-derived open reading frames (ORFs) to investigate the functional potential of short open reading frames (sORFs), which encode proteins less than 100 amino acids long (microproteins). By overlaying statistically significant single-nucleotide variants (SNVs) from the GWAS Catalog with a set of sORFs supported by multiple Ribo-seq studies, we highlight previously unannotated genomic regions with potential coding functionality. This strategy resulted in a shortlist of six high-priority candidates, each harboring disease- or trait-associated SNVs within microprotein-encoding regions. These candidates are strong contenders for downstream functional validation via gene editing, overexpression, or interaction assays in a wet lab setting.

Supported by the high Combined Annotation Dependent Depletion (CADD) scores, these SNVs are highly likely to be biologically meaningful and potentially alter the structure or function of the predicted microproteins. We excluded potentially false-positive SNVs that affect already annotated genes to prevent attributing known gene functions to predicted microproteins. However, even true overlaps may be complex regulatory or dual-function loci, which happen to be located within studied sORFs.

This approach demonstrates that large-scale genetic association data can be leveraged to uncover protein-coding potential in previously unannotated genomic regions. We therefore provide a practical framework for prioritizing novel proteins for the subsequent experimental validation by integrating population-level variant data with advances in human genome annotation. However, several limitations of the work must be acknowledged. First, to remove complexity in this exploratory methodology, we excluded sORFs that contain intronic regions; for more comprehensive analysis, more advanced bioinformatic computations are needed to fully capture and interpret such transcriptomic features. Second, by retaining only sORFs identified by at least two independent Ribo-seq studies, we increase the reliability of our candidate list but may have inadvertently excluded functional microproteins that are tissue-specific or expressed under specific conditions. Finally, protein structure predictions by AlphaFold are not well-defined for some of the selected candidates, and such microproteins

may still exert biological functions. Disordered protein domains could participate in regulatory processes, including protein-protein interactions, subcellular localization signalling, often contain protein cleavage motifs, or can possess HLA-binding properties.

Overall, these results reinforce the growing consensus that microproteins, despite being historically neglected, may serve essential biological functions and merit inclusion in functional genome annotations. If experimentally validated, the microproteins identified in this study could form a new class of biomolecules involved in human disease, offering potential usage as biomarkers or therapeutic targets.

SUMMARY

Short open reading frames (sORFs) that encode microproteins remain largely missing from current human genome annotations. This study aimed to investigate their potential functional relevance by leveraging single-nucleotide variant (SNV) associations with common human traits and diseases, as catalogued in the GWAS Catalog.

In this work, we focused on sORFs identified in prior ribosome profiling (Ribo-seq) studies and selected SNVs located within these regions. After prioritizing the variants based on their CADD scores, we identified six sORFs as strong candidates for experimental follow-up.

Overall, this work demonstrates the value of integrating large-scale genetic association data with emerging annotations to uncover overlooked coding elements and provides a framework for future studies on microprotein function.

REFERENCES

Aigner, J., Villatoro, S., Rabionet, R., Roquer, J., Jiménez-Conde, J., Martí, E., & Estivill, X. (2013). A common 56-kilobase deletion in a primate-specific segmental duplication creates a novel butyrophilin-like protein. *BMC Genetics*, *14*. https://doi.org/10.1186/1471-2156-14-61

Auton, A., Abecasis, G. R., Altshuler, D. M., Durbin, R. M., Bentley, D. R., Chakravarti, A., Clark, A. G., Donnelly, P., Eichler, E. E., Flicek, P., Gabriel, S. B., Gibbs, R. A., Green, E. D., Hurles, M. E., Knoppers, B. M., Korbel, J. O., Lander, E. S., Lee, C., Lehrach, H., … Schloss, J. A. (2015). A global reference for human genetic variation. In *Nature* (Vol. 526, Issue 7571). https://doi.org/10.1038/nature15393

Basrai, M. A., Hieter, P., & Boeke, J. D. (1997). Small open reading frames: Beautiful needles in the haystack. In *Genome Research* (Vol. 7, Issue 8). https://doi.org/10.1101/gr.7.8.768

Calviello, L., Mukherjee, N., Wyler, E., Zauber, H., Hirsekorn, A., Selbach, M., Landthaler, M., Obermayer, B., & Ohler, U. (2016). Detecting actively translated open reading frames in ribosome profiling data. *Nature Methods*, *13*(2), 165–170. https://doi.org/10.1038/nmeth.3688

Cerezo, M., Sollis, E., Ji, Y., Lewis, E., Abid, A., Bircan, K. O., Hall, P., Hayhurst, J., John, S., Mosaku, A., Ramachandran, S., Foreman, A., Ibrahim, A., McLaughlin, J., Pendlington, Z., Stefancsik, R., Lambert, S. A., McMahon, A., Morales, J., Keane, T., … Harris, L. W. (2025). The NHGRI-EBI GWAS Catalog: standards for reusability, sustainability and diversity. *Nucleic acids research*, *53*(D1), D998–D1005. https://doi.org/10.1093/nar/gkae1070

Chen, J., Brunner, A.-D., Cogan, J. Z., Nuñez, J. K., Fields, A. P., Adamson, B., Itzhak, D. N., Li, J. Y., Mann, M., Leonetti, M. D., & Weissman, J. S. (2020). Pervasive functional translation of noncanonical human open reading frames. *Science*, *367*(6482), 1140–1146. https://doi.org/10.1126/science.aay0262

Collins, F. S., Brooks, L. D., & Chakravarti, A. (1998). A DNA polymorphism discovery resource for research on human genetic variation. *Genome Research*, *8*(12). https://doi.org/10.1101/gr.8.12.1229

Couso, J. P., & Patraquim, P. (2017). Classification and function of small open reading frames. In *Nature Reviews Molecular Cell Biology* (Vol. 18, Issue 9). https://doi.org/10.1038/nrm.2017.58

Deutsch, E. W., Kok, L. W., Mudge, J. M., Ruiz-Orera, J., Fierro-Monti, I., Sun, Z., Abelin, J. G., Alba, M. M., Aspden, J. L., Bazzini, A. A., Bruford, E. A., Brunet, M. A., Calviello, L., Carr, S. A., Carvunis, A.-R., Chothani, S., Clauwaert, J., Dean, K., Faridi, P., Frankish, A., ... van Heesch, S. (2024). High-quality peptide evidence for annotating non-canonical open reading frames as human proteins. *bioRxiv*. https://doi.org/10.1101/2024.09.09.612016

Frith, M. C., Forrest, A. R., Nourbakhsh, E., Pang, K. C., Kai, C., Kawai, J., Carninci, P., Hayashizaki, Y., Bailey, T. L., & Grimmond, S. M. (2006). The abundance of short proteins in the mammalian proteome. *PLoS Genetics*, *2*(4). https://doi.org/10.1371/journal.pgen.0020052

Gaertner, B., Van Heesch, S., Schneider-Lunitz, V., Schulz, J. F., Witte, F., Blachut, S., Nguyen, S., Wong, R., Matta, I., Hübner, N., & Sander, M. (2020). A human ESC-based screen identifies a role for the translated lncRNA LINC00261 in pancreatic endocrine differentiation. *eLife*, *9*, e58659. https://doi.org/10.7554/eLife.58659

Henikoff, S., & Henikoff, J. G. (1992). Amino acid substitution matrices from protein blocks. *Proceedings of the National Academy of Sciences of the United States of America*, *89*(22). https://doi.org/10.1073/pnas.89.22.10915

Hollox, E. J., Zuccherato, L. W., & Tucci, S. (2022). Genome structural variation in human evolution. In *Trends in Genetics* (Vol. 38, Issue 1). https://doi.org/10.1016/j.tig.2021.06.015

Ingolia, N. T. (2016). Ribosome Footprint Profiling of Translation throughout the Genome. In *Cell* (Vol. 165, Issue 1). https://doi.org/10.1016/j.cell.2016.02.066

Ingolia, N. T., Ghaemmaghami, S., Newman, J. R. S., & Weissman, J. S. (2009). Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling. *Science*, *324*(5924). https://doi.org/10.1126/science.1168978

Ji, Z., Song, R., Regev, A., & Struhl, K. (2015). Many lncRNAs, 5'UTRs, and pseudogenes are translated and some are likely to express functional proteins. *eLife*, *4*, e08890. https://doi.org/10.7554/eLife.08890

Kaushik, S., Kaushik, S., & Sharma, D. (2018). Functional genomics. In *Encyclopedia of Bioinformatics and Computational Biology: ABC of Bioinformatics* (Vols. 1–3). https://doi.org/10.1016/B978-0-12-809633-8.20222-7

Ladoukakis, E., Pereira, V., Magny, E. G., Eyre-Walker, A., & Couso, J. P. (2011). Hundreds of putatively functional small open reading frames in Drosophila. *Genome Biology*, *12*(11). https://doi.org/10.1186/gb-2011-12-11-r118

Leong, A. Z. X., Lee, P. Y., Mohtar, M. A., Syafruddin, S. E., Pung, Y. F., & Low, T. Y. (2022). Short open reading frames (sORFs) and microproteins: an update on their identification and validation measures. In *Journal of Biomedical Science* (Vol. 29, Issue 1). https://doi.org/10.1186/s12929-022-00802-5

Levy, S., Sutton, G., Ng, P. C., Feuk, L., Halpern, A. L., Walenz, B. P., Axelrod, N., Huang, J., Kirkness, E. F., Denisov, G., Lin, Y., MacDonald, J. R., Pang, A. W. C., Shago, M., Stockwell, T. B., Tsiamouri, A., Bafna, V., Bansal, V., Kravitz, S. A., … Venter, J. C. (2007). The diploid genome sequence of an individual human. *PLoS Biology*, *5*(10). https://doi.org/10.1371/journal.pbio.0050254

Makarewich, C. A., & Olson, E. N. (2017). Mining for Micropeptides. In *Trends in Cell Biology* (Vol. 27, Issue 9). https://doi.org/10.1016/j.tcb.2017.04.006

Marees, A. T., de Kluiver, H., Stringer, S., Vorspan, F., Curis, E., Marie-Claire, C., & Derks, E. M. (2018). A tutorial on conducting genome-wide association studies: Quality control and statistical analysis. *International Journal of Methods in Psychiatric Research*, *27*(2). https://doi.org/10.1002/mpr.1608

Martinez, T. F., Chu, Q., Donaldson, C., Tan, D., Shokhirev, M. N., & Saghatelian, A. (2020). Accurate annotation of human protein-coding small open reading frames. *Nature Chemical Biology*, *16*(4), 458–468. https://doi.org/10.1038/s41589-019-0425-0

McClellan, J., & King, M. C. (2010). Genetic heterogeneity in human disease. In *Cell* (Vol. 141, Issue 2). https://doi.org/10.1016/j.cell.2010.03.032

Montavon, T., Thevenet, L., & Duboule, D. (2012). Impact of copy number variations (CNVs) on long-range gene regulation at the HoxD locus. *Proceedings of the National Academy of Sciences of the United States of America*, *109*(50). https://doi.org/10.1073/pnas.1217659109

Mudge, J. M., Ruiz-Orera, J., Prensner, J. R., Brunet, M. A., Calvet, F., Jungreis, I., Gonzalez, J. M., Magrane, M., Martinez, T. F., Schulz, J. F., Yang, Y. T., Albà, M. M., Aspden, J. L., Baranov, P. V., Bazzini, A. A., Bruford, E., Martin, M. J., Calviello, L., Carvunis, A.-R., … Van Heesch, S. (2022). Standardized annotation of translated open reading frames. *Nature Biotechnology*, *40*(7), 994–999. https://doi.org/10.1038/s41587-022-01369-0

Olexiouk, V., van Criekinge, W., & Menschaert, G. (2018). An update on sORFs.org: A repository of small ORFs identified by ribosome profiling. *Nucleic Acids Research*, *46*(D1). https://doi.org/10.1093/nar/gkx1130

Raj, A., Wang, S. H., Shim, H., Harpak, A., Li, Y. I., Engelmann, B., Stephens, M., Gilad, Y., & Pritchard, J. K. (2016). Thousands of novel translated open reading frames in humans inferred by ribosome footprint profiling. *eLife*, *5*, e13328. https://doi.org/10.7554/eLife.1332

Saghatelian, A., & Couso, J. P. (2015). Discovery and characterization of smORF-encoded bioactive polypeptides. In *Nature Chemical Biology* (Vol. 11, Issue 12). https://doi.org/10.1038/nchembio.1964

Schlesinger, D., & Elsässer, S. J. (2022). Revisiting sORFs: overcoming challenges to identify and characterize functional microproteins. In *FEBS Journal* (Vol. 289, Issue 1). https://doi.org/10.1111/febs.15769

Sharp, A. J., Cheng, Z., & Eichler, E. E. (2006). Structural variation of the human genome. In *Annual Review of Genomics and Human Genetics* (Vol. 7). https://doi.org/10.1146/annurev.genom.7.080505.115618

Sieber, P., Platzer, M., & Schuster, S. (2018). The Definition of Open Reading Frame Revisited. In *Trends in Genetics* (Vol. 34, Issue 3). https://doi.org/10.1016/j.tig.2017.12.009

Taliun, D., Harris, D. N., Kessler, M. D., Carlson, J., Szpiech, Z. A., Torres, R., Taliun, S. A. G., Corvelo, A., Gogarten, S. M., Kang, H. M., Pitsillides, A. N., LeFaive, J., Lee, S. been,

Tian, X., Browning, B. L., Das, S., Emde, A. K., Clarke, W. E., Loesch, D. P., … Abecasis, G. R. (2021). Sequencing of 53,831 diverse genomes from the NHLBI TOPMed Program. *Nature*, *590*(7845). https://doi.org/10.1038/s41586-021-03205-y

Uffelmann, E., Huang, Q. Q., Munung, N. S., De Vries, J., Okada, Y., Martin, A. R., Martin, H. C., Lappalainen, T., & Posthuma, D. (2021). Genome-wide association studies. *Nature Reviews Methods Primers*, *1*(1), 59. https://doi.org/10.1038/s43586-021-00056-9

Van Heesch, S., Witte, F., Schneider-Lunitz, V., Schulz, J. F., Adami, E., Faber, A. B., Kirchner, M., Maatz, H., Blachut, S., Sandmann, C.-L., Kanda, M., Worth, C. L., Schafer, S., Calviello, L., Merriott, R., Patone, G., Hummel, O., Wyler, E., Obermayer, B., … Hubner, N. (2019). The Translational Landscape of the Human Heart. *Cell*, *178*(1), 242-260.e29. https://doi.org/10.1016/j.cell.2019.05.010

Wang, Z., & Moult, J. (2001). SNPs, protein structure, and disease. *Human Mutation*, *17*(4). https://doi.org/10.1002/humu.22

Wheeler, D. A., Srinivasan, M., Egholm, M., Shen, Y., Chen, L., McGuire, A., He, W., Chen, Y. J., Makhijani, V., Roth, G. T., Gomes, X., Tartaro, K., Niazi, F., Turcotte, C. L., Irzyk, G. P., Lupski, J. R., Chinault, C., Song, X. Z., Liu, Y., … Rothberg, J. M. (2008). The complete genome of an individual by massively parallel DNA sequencing. *Nature*, *452*(7189). https://doi.org/10.1038/nature06884

SUPPLEMENTARY

The repository below contains code and datasets relevant to this work.

https://github.com/svyat1kk/microproteins

Non-exclusive licence to reproduce the thesis and make the thesis public

I, Sviatoslav-Oleh Savchak

1. grant the University of Tartu a free permit (non-exclusive licence) to reproduce, for the purpose of preservation, including for adding to the digital archives of the University of Tartu until the expiry of the term of copyright, my thesis "Identifying microproteins with genetic association data" supervised by Anastasiia Alekseienko and Erik Abner
2. grant the University of Tartu a permit to make the thesis specified in point 1 available to the public via the web environment of the University of Tartu, including via the digital archives, under the Creative Commons licence CC BY NC ND 4.0, which allows, by giving appropriate credit to the author, to reproduce, distribute the work and communicate it to the public, and prohibits the creation of derivative works and any commercial use of the work until the expiry of the term of copyright;
3. am aware of the fact that the author retains the rights specified in points 1 and 2;
4. confirm that granting the non-exclusive licence does not infringe other persons' intellectual property rights or rights arising from the personal data protection legislation.

Sviatoslav-Oleh Savchak
20/05/2025