Bioinformatic Discovery of Clinically Relevant Splicing Variants in Public Datasets

Rugare Maruzani, r.maruzani@liverpool.ac.uk, University of Liverpool

Supervisor: Dr Ali Awan, ali.awan@kcl.ac.uk, Genomics Innovation Unit, Guy's Hospital

Introduction

- **University of Liverpool** – MRC DiMeN DTP Computational Biology
- My PhD focuses on developing tools to better detect low frequency ctDNA variants
- University of Sheffield - MSc in Molecular Medicine
- My interests are in leveraging computational tools to improve human health
- This project was an opportunity to work on a project that could contribute to improving health using computational methods

Graduating from Molecular Medicine University of Sheffield, 2019

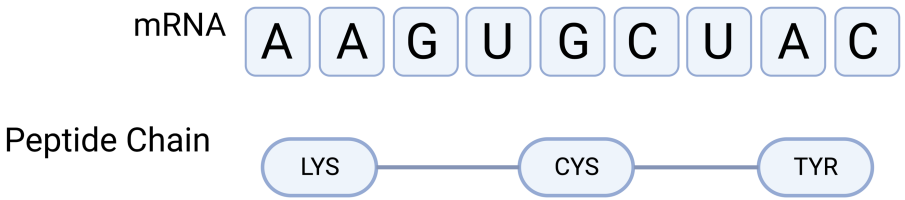
How Did I End Up Here?

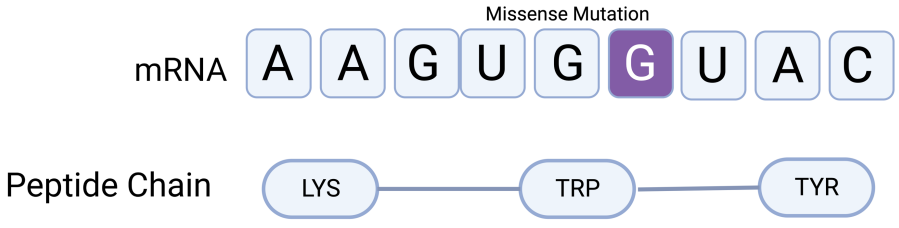
- MRC and DiMeN are supportive of PhD student internships
- DiMeN has an extensive network of former and current students
- Today, I will present the work I have been involved with over the last three months
- Ideas and experimental approaches of the project
- Outcomes

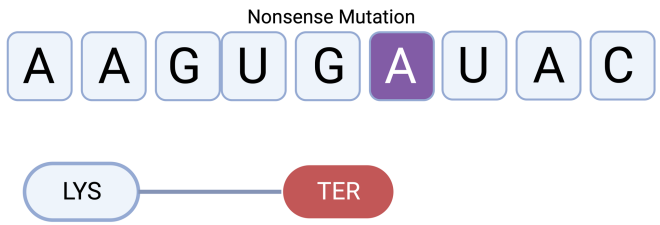
Dr Jack Paveley
Business Development Manager
APIS Assay Technologies Ltd.

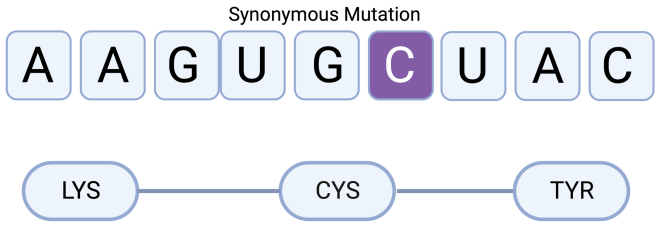
Genetics Testing and Rare Diseases

- Approximately 80% of ‘rare’ diseases have a genetic component
- Genetic testing is routinely used to try and diagnosis rare diseases
- Synonymous variants are often overlooked compared to variants with obvious consequences on protein production





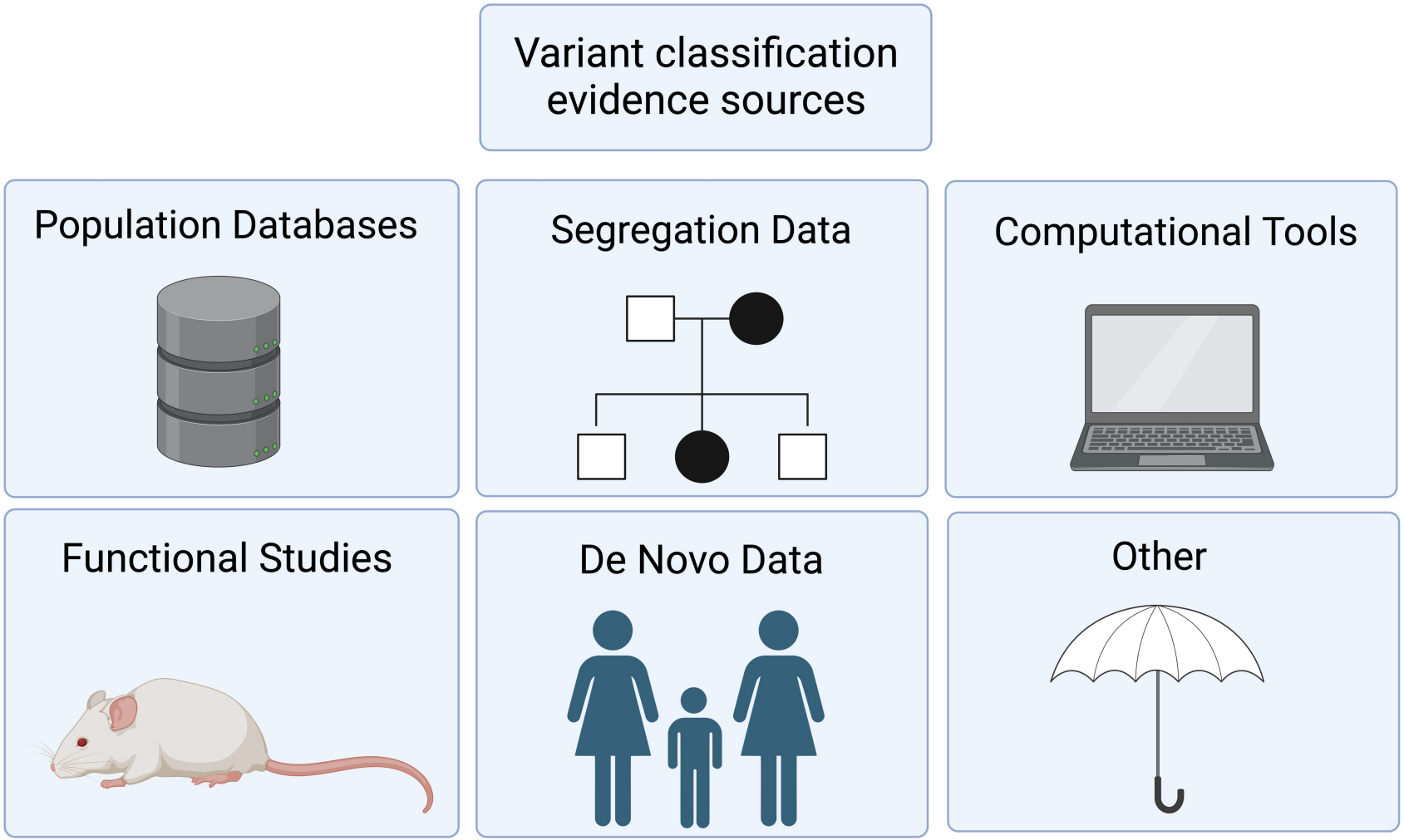




ClinVar is A Database of Variants and Their Relationship to Human Disease

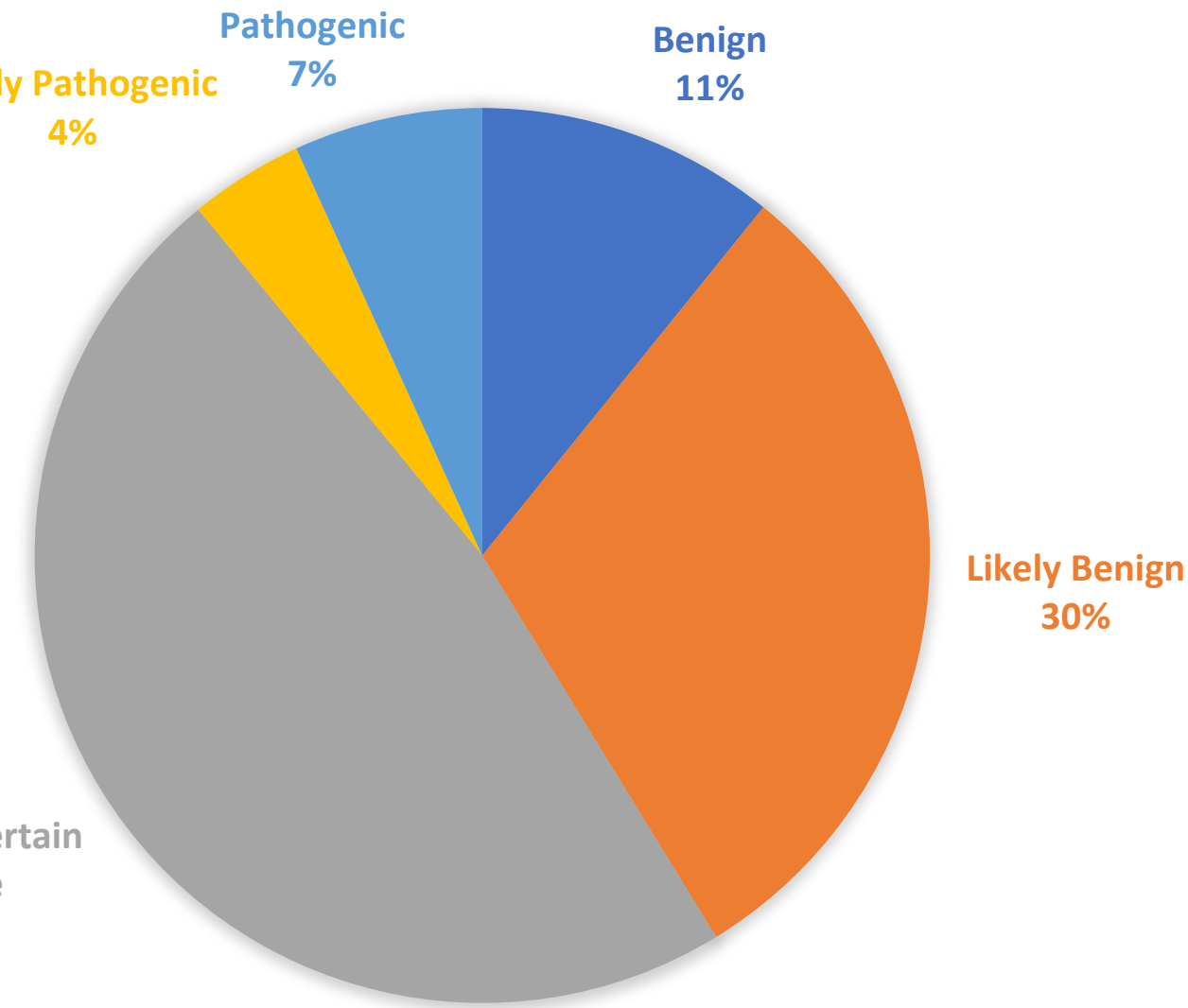
- Variants are classified into one of 5 categories based on clinical significance
    I. Benign
    II. Likely Benign
    III. Variant of Uncertain Significance (VUS)
    IV. Likely Pathogenic
    V. Pathogenic

Evidence Sources for Clinical Significance Classification

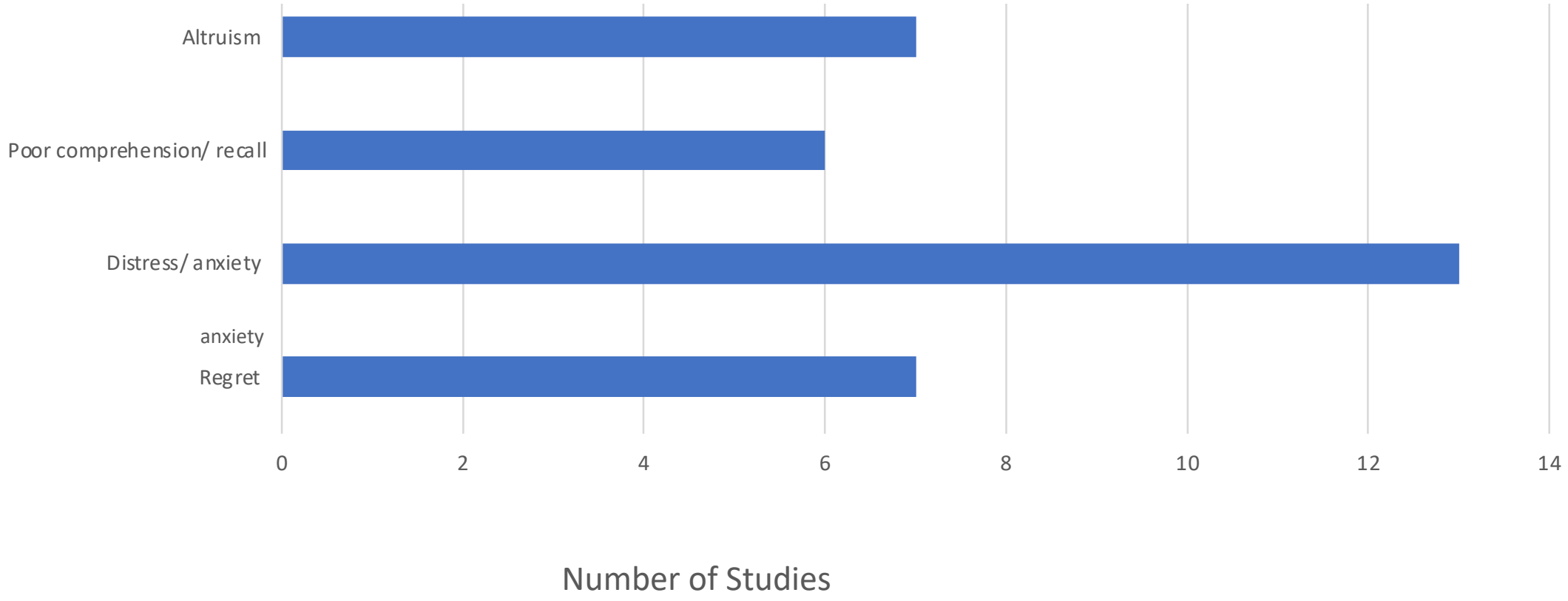


The Five ClinVar Variant Classifications

n = 2,179,119



VUSs Cause Patient Distress



Clift, K., Macklin, S., Halverson, C. *et al.* Patients' views on variants of uncertain significance across indications. *J Community Genet* **11**, 139–145 (2020). https://doi.org/10.1007/s12687-019-00434-7

Reclassifying VUSs May Improve Prospects for Rare Disease Diagnosis

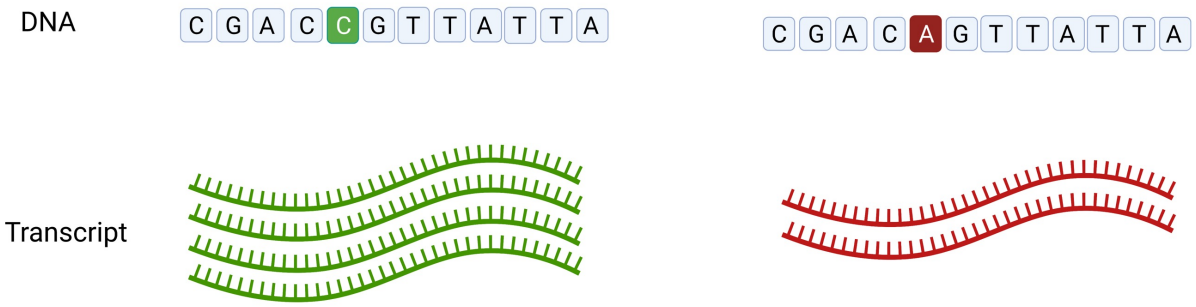
- Resolving VUSs may hold information that can help diagnose rare diseases
- VUSs are more prevalent in non-European ancestry individuals, with consequences for health inequality

Project Aim

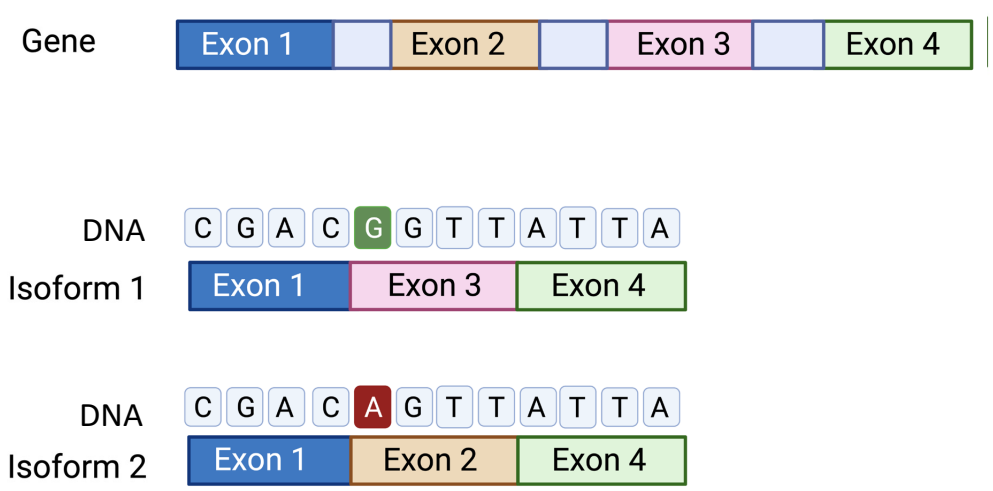
Develop a computational pipeline to characterise the effects that VUSs have on transcript splicing isoforms and expression levels

Overview of Research Questions

Which VUSs (if any) affect expression?



Which VUSs (if any) affect splicing



The Sequence Read Archive

- SRA is the largest (mostly) public repository of NGS data, upwards of 17 Petabytes of data – **1 Petabytes** = **1 million Gb**
- The database contains both DNA and RNA sequencing data from mostly human and mouse
- We are searching RNA sequencing data
- Querying the SRA for samples with a particular sequence is a huge challenge due to the size of the database

Genomes | Genome Browser | Tools | Mirrors | Downloads | My Data | Projects | Help | About Us

Human BLAT Search

BLAT Search Genome

Genome: ☐ Search all genomes

Human

Assembly:

Dec. 2013 (GRCh38/hg38)

Query type:

BLAT's guess

Sort output:

query,score

Output type:

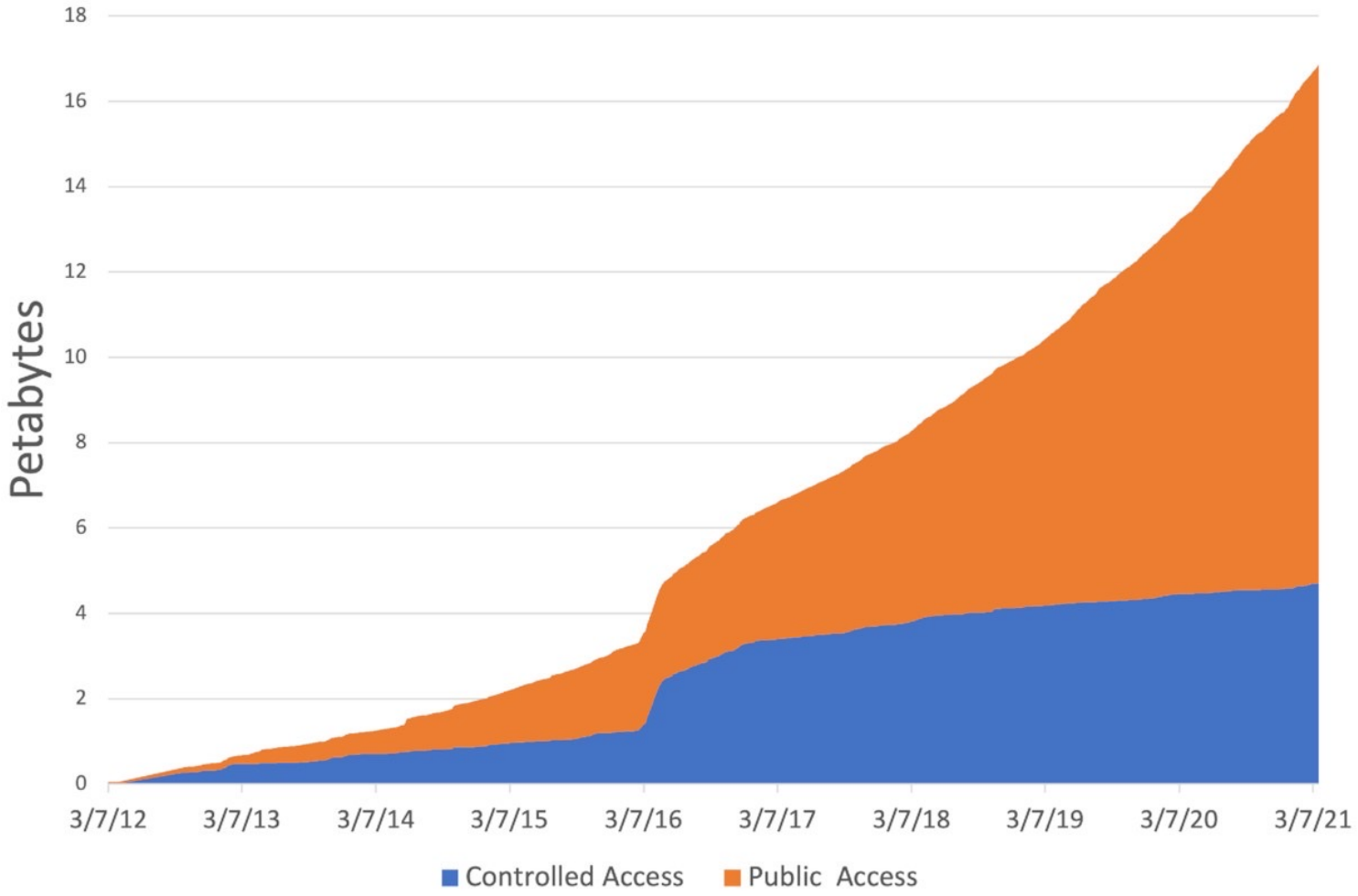
hyperlink

GTCCTCGGAACCAGGACCTCGGCGTGGCCTAGCG

☐ All Results (no minimum matches)

Submit | I'm feeling lucky | Clear

The SRA Explosive Growth



Kenneth Katz and others, **The Sequence Read Archive: a decade more of explosive growth**, *Nucleic Acids Research*, Volume 50, Issue D1, 7 January 2022, Pages D387–D390, https://doi.org/10.1093/nar/gkab1053

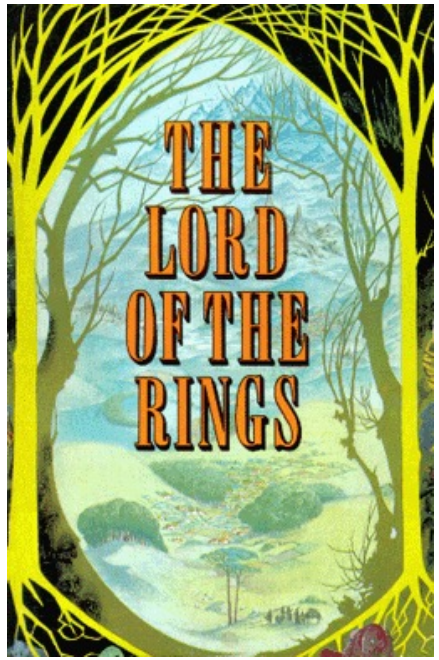
The Querying Problem

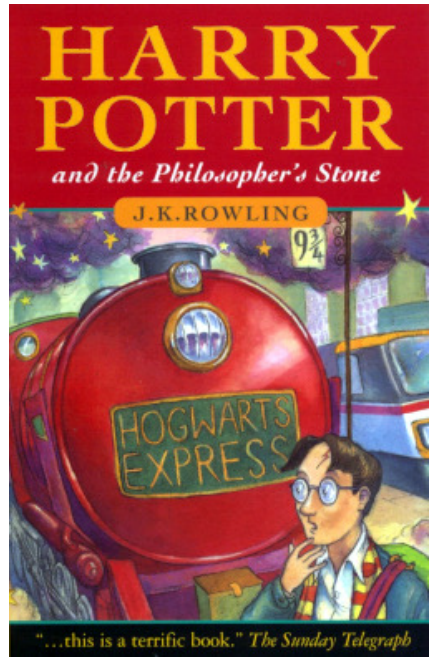
Which book does each sentence appear in? What page?

"Harry - yer a wizard."

"The road goes ever on and on."

"In a hole in the ground there lived a hobbit."





Index

...

“Harry - yer a wizard.”, **Harry Potter and the Philosopher’s Stone**, Page 92

....

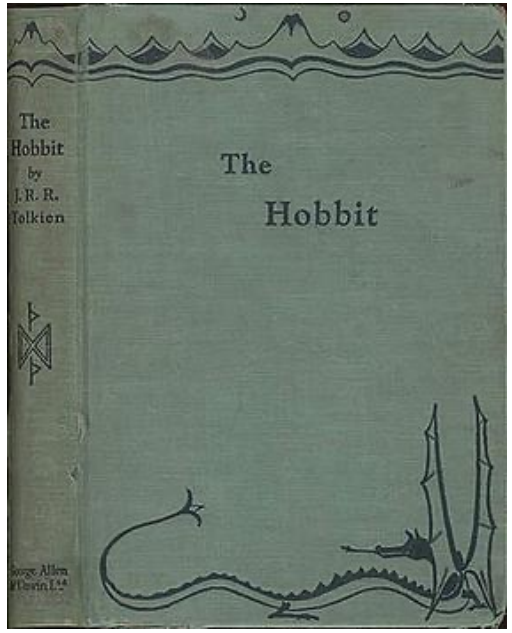
....

“In a hole in the ground there lived a hobbit.”, **The Hobbit,** Page 8

....

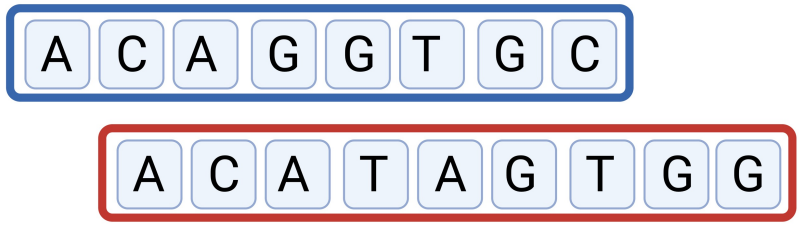
"The road goes ever on and on.”, **The Lord of The Rings**, Page 34

...



Bifrost Enables Indexing of RNA-Seq Data

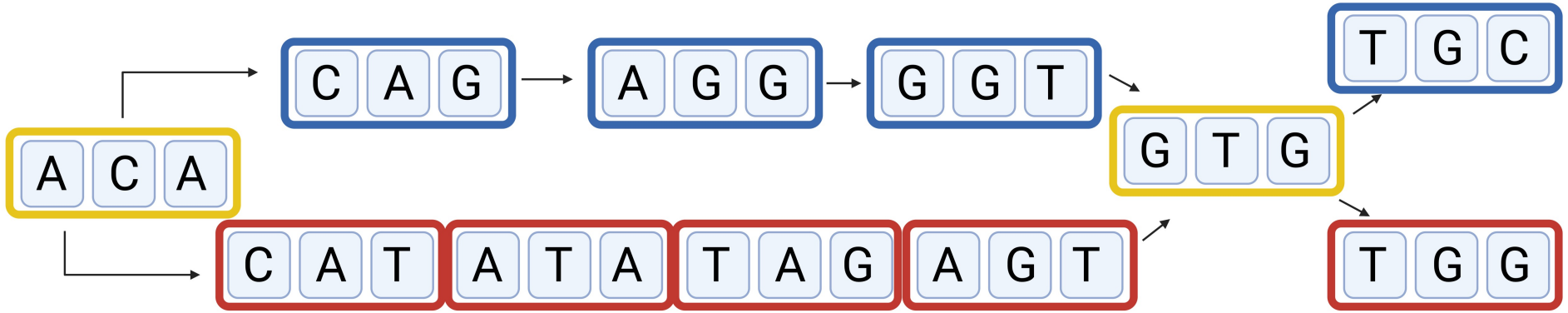
FASTQ files



FASTQ File Size = 1 Tb
Bifrost Index Size = 60 Gb

kmer = 3

Coloured de Bruijn graph



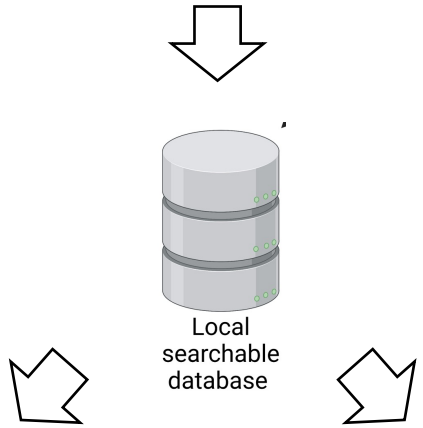
Querying Bifrost Index

- Bifrost takes FASTA files as query input
- Bifrost can query the index and return FASTQ files that contain the query sequence
- Bifrost returns the original RNA-Seq sample containing the query sequence

Pipeline Workflow

Variant ID	REF	ALT	Gene	Ref Q-Sequence	Alt Q-Sequence
1552	A	G	TP53	ATGTATCGGATTACG**(A)**ACTTCACGGAGACTA	ATGTATCGGATTACG**(G)**ACTTCACGGAGACTA
…	…	…	…	…	…

ClinVar variants = 1233, VUS, synonymous, SNPs



Reference Q-Sequence

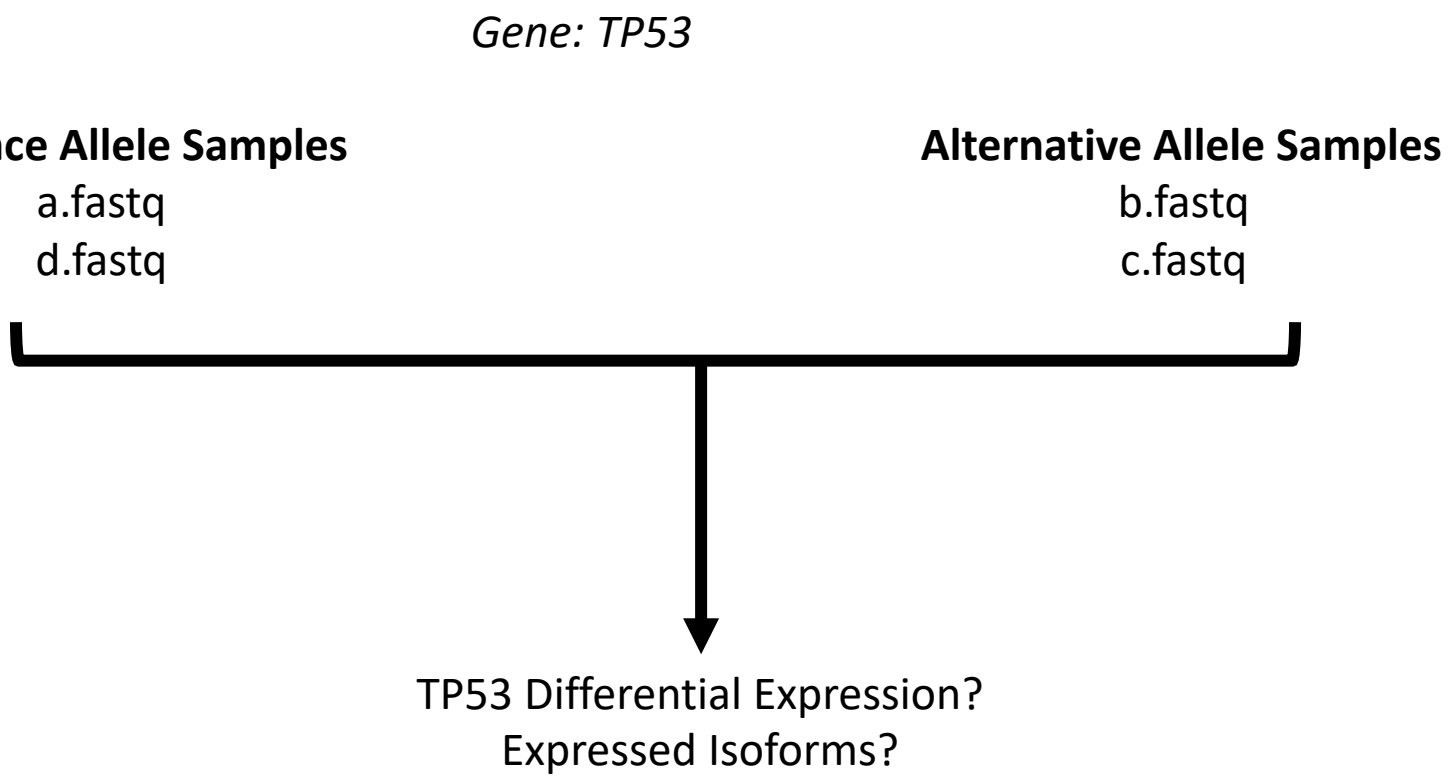
Variant ID	a.fastq	b.fastq	c.fastq	d.fastq
1552	1	0	0	1
…	…	…	…	…
…	…	…	…	…

Alternate Q-Sequence

Variant ID	a.fastq	b.fastq	c.fastq	d.fastq
1552	0	1	1	0
…	…	…	…	…
…	…	…	…	…

Pipeline Workflow

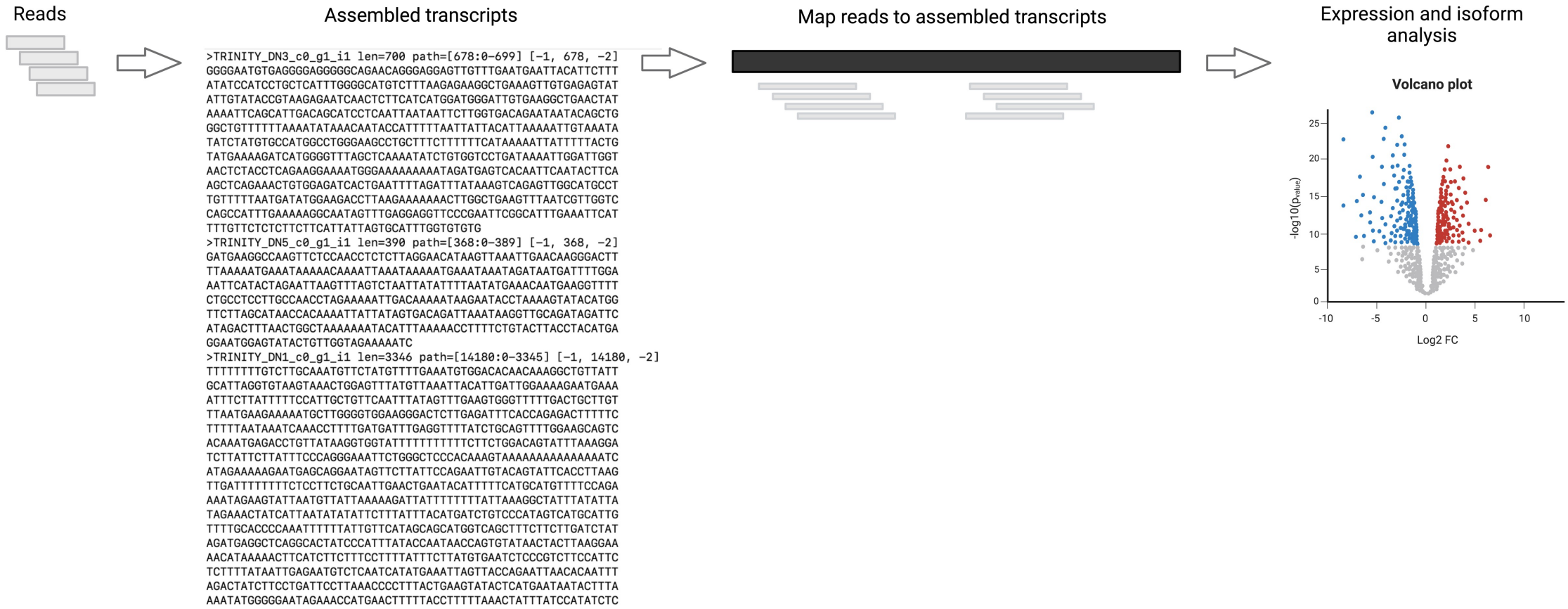
Variant ID: 1552



Quantifying Transcript Expression Trinity

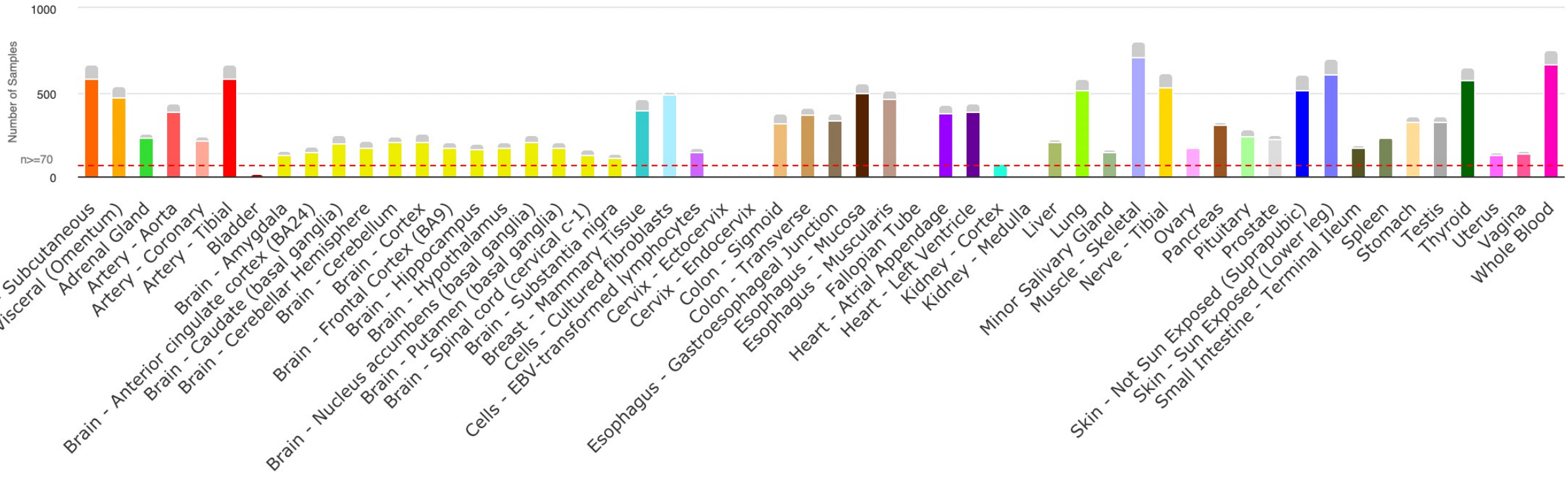
- Trinity assembles transcript sequences from RNA-Seq data

Gene only reads*



Stretch Goals: Tissue Specific Effects

- **GTEx Database** holds tissue-specific gene expression and regulation data



https://gtexportal.org/home/tissueSummaryPage

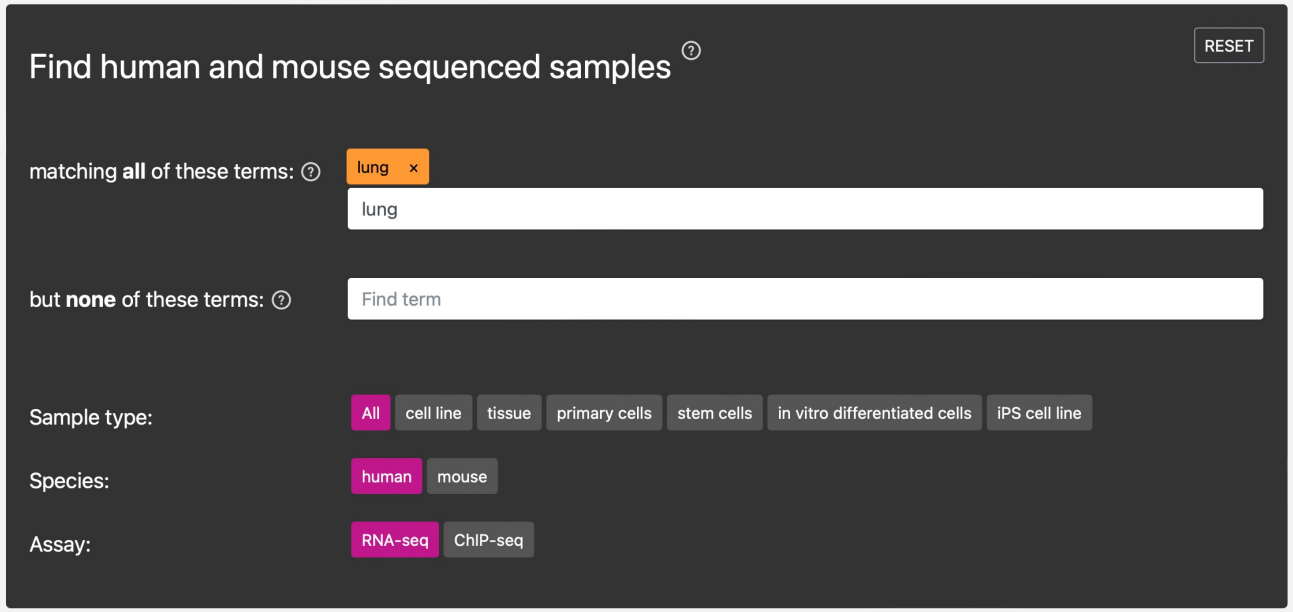
Stretch Goals: Tissue Specific Effects

- **MetaSRA** provides a workaround to access tissue specific samples

24 MetaSRA tissues – Appx. 100 Gb per tissue

MetaSRA

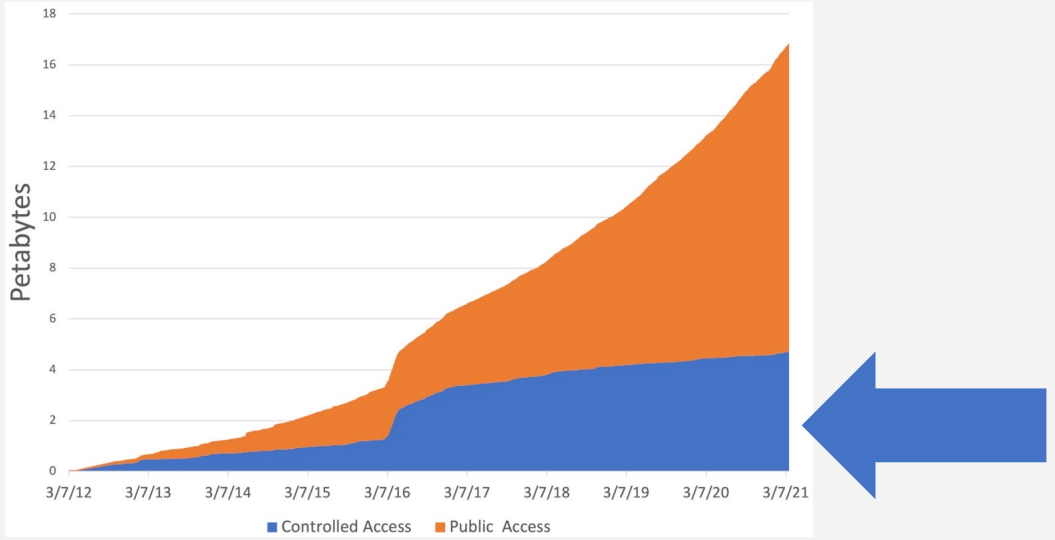
Normalized metadata for the Sequence Read Archive



Found **16701** samples in **464** studies

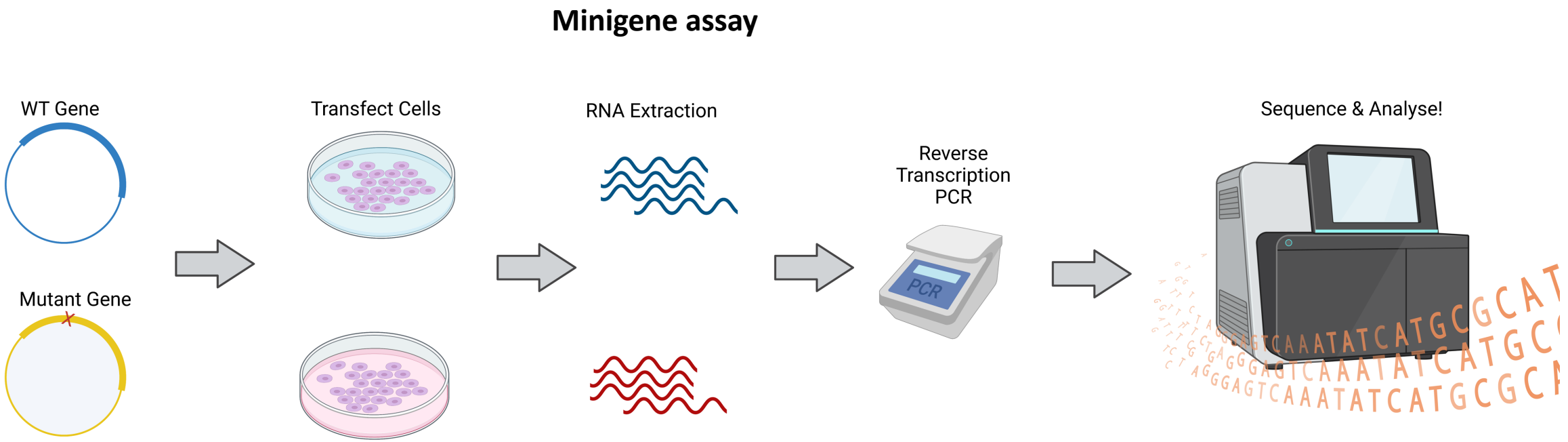
Download

```
rmaruzani@PCA100102:~/storage/hydra2/vus/toolkits/bifrost/parallel/whole_studies$
DRP003703-COLON-DRR078748.fastq   SRP058414-LYMPH-SRR2029642.fastq
DRP003703-COLON-DRR078749.fastq   SRP058414-LYMPH-SRR2029643.fastq
DRP003703-COLON-DRR078750.fastq   SRP058414-LYMPH-SRR2029644.fastq
DRP003703-COLON-DRR078751.fastq   SRP058414-LYMPH-SRR2031977.fastq
DRP003703-COLON-DRR078752.fastq   SRP058414-LYMPH-SRR2031990.fastq
DRP003703-COLON-DRR078753.fastq   SRP058414-LYMPH-SRR2032007.fastq
DRP003703-COLON-DRR078754.fastq   SRP058414-LYMPH-SRR2032020.fastq
DRP003703-COLON-DRR078755.fastq   SRP058414-LYMPH-SRR2032037.fastq
DRP003703-COLON-DRR078756.fastq   SRP058414-LYMPH-SRR2032038.fastq
DRP003703-COLON-DRR078757.fastq   SRP058414-LYMPH-SRR2032039.fastq
DRP003703-COLON-DRR078758.fastq   SRP058414-LYMPH-SRR2032041.fastq
DRP003703-COLON-DRR078759.fastq   SRP058414-LYMPH-SRR2032042.fastq
DRP003703-COLON-DRR078760.fastq   SRP058414-LYMPH-SRR2032043.fastq
```



Stretch Goals: Experimental Validation

- **If promising variants are discovered** – perform experimental validation



Houston, we have a problem...

Quality control rejects most variants

In samples where Bifrost finds a VUS:

1) Filter out all reads not mapping to the relevant gene > is the query sequence still still there?

2) For each VUS, we should have at least 5 samples supporting reference and alternate alleles

- **1233 synonymous VUS reduced to 15** we can assess for differences in expression and splice patterns

This approach is not practical for reclassifying VUS

- Requires a huge amount SRA data in local database to get a handful of variants to assess

Indexing NGS data is computationally expensive

With unlimited (to Bifrost) CPU and RAM

- Building a Bifrost index with 1 Tb of FASTQ files = **3 days**
- Building a Bifrost index with 2 Tb of FASTQ files = **Nearly 7 days!**

Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences

Alexander Rives[a,b,1,2], Joshua Meier[a,1], Tom Sercu[a,1], Siddharth Goyal[a,1], Zeming Lin[b], Jason Liu[a], Demi Guo[c,3], Myle Ott[a], C. Lawrence Zitnick[a], Jerry Ma[d,e,3], and Rob Fergus[b]

[a]Facebook AI Research, New York, NY 10003; [b]Department of Computer Science, New York University, New York, NY 10012; [c]Harvard University, Cambridge, MA 02138; [d]Booth School of Business, University of Chicago, Chicago, IL 60637; and [e]Yale Law School, New Haven, CT 06511

Edited by David T. Jones, University College London, London, United Kingdom, and accepted by Editorial Board Member William H. Press December 16, 2020 (received for review August 6, 2020)

In the field of artificial intelligence, a combination of scale in data and model capacity enabled by unsupervised learning has led to major advances in representation learning and statistical generation. In the life sciences, the anticipated growth of sequencing promises unprecedented data on natural sequence diversity. Protein language modeling at the scale of evolution is a logical step toward predictive and generative artificial intelligence for biology. To this end, we use unsupervised learning to train a deep contextual

Prediction of Mutational Effects

The mutational fitness landscape provides deep insight into biology. Coupling next-generation sequencing with a mutagenesis screen allows parallel readout of tens of thousands of variants of a single protein (62). The detail and coverage of these experiments provides a view into the mutational fitness landscape of individual proteins, giving quantitative relationships between sequence and protein function. We adapt the Transformer protein language model to predict the quantitative effect of mutations.

First, we investigate intraprotein variant effect prediction, where a limited sampling of mutations is used to predict the effect of unobserved mutations. This setting has utility in protein engineering applications (63). We evaluate the representations on two deep mutational scanning datasets used by recent state-of-the-art methods for variant effect prediction, Envision (64) and DeepSequence (26). Collectively, the data includes over 700,000 variant effect measurements from over 100 large-scale experimental mutagenesis datasets.

Conclusions

- Resolving VUSs can contribute to the diagnosis of rare diseases
- We attempted to develope a pipeline that takes advantage of the SRA database to resolve VUSs
- A challenge of working with the database is querying a given sequence in a large database
- Ultimately, due to the large amounts of data and compute resources required – this approach is not practical for reclassifying VUS variants
- Other methods are emerging for predicting mutation effects which could be used to reclassify VUSs

Genomics Innovation Unit

Dr Ali Awan

PhD Supervisory Team

Dr Anna Fowler

Prof Andrea Jorgensen

Dr Liam Brierley

and others

UKRI Medical Research Council