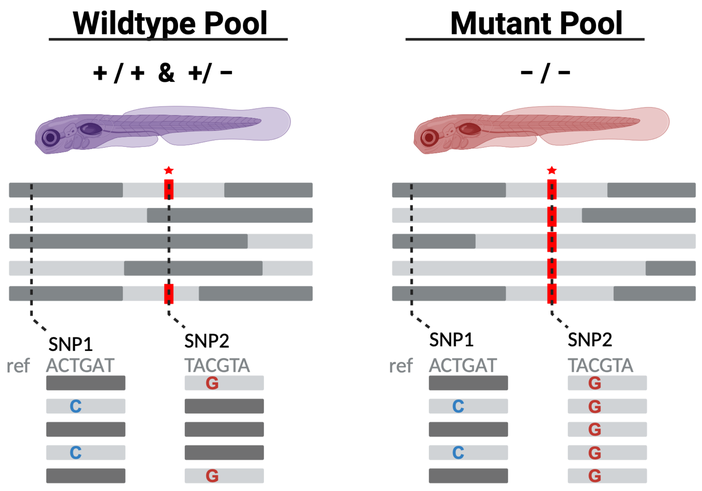
# Bioinformatic Tool for Identifying Causative Mutation Candidates from RNA-seq Data

Ramzy Al-Mulla[1], Amelia Dayton[1], Zach Girard[1], Anne Martin[2], Adam Miller[2]

1. Bioinformatics and Genomics Master's Program - KCGIP, University of Oregon; 2. Institute of Neuroscience, Department of Biology, University of Oregon

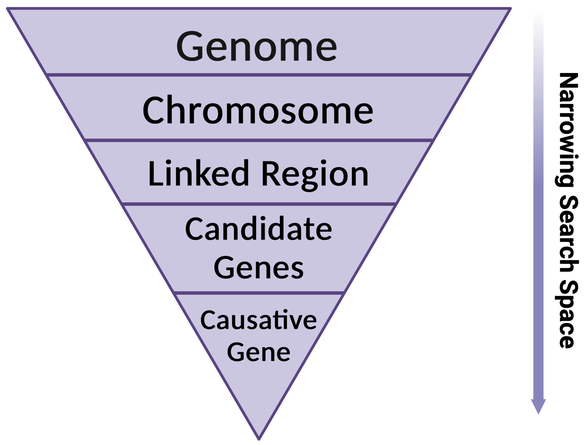
## 1. How do we identify novel causative mutations?

Bulk Segregant Analysis (BSA) is a method used to identify causative mutations and examine related gene regulatory changes in a mutant population.




However, identifying an individual mutation within the genome is largely inaccessible to researchers without computational expertise. Existing tools[1,2] are outdated and rely on unsupported software.

**Objective:** Build a fast, accessible bioinformatics pipeline to identify causative mutations in the genome




With our new pipeline, we aim to uncover an unknown mutation that resulted in the diminished development of chemical and electrical synapses in zebrafish.

## 2. Bioinformatic workflow from RNA-seq data to candidate genes

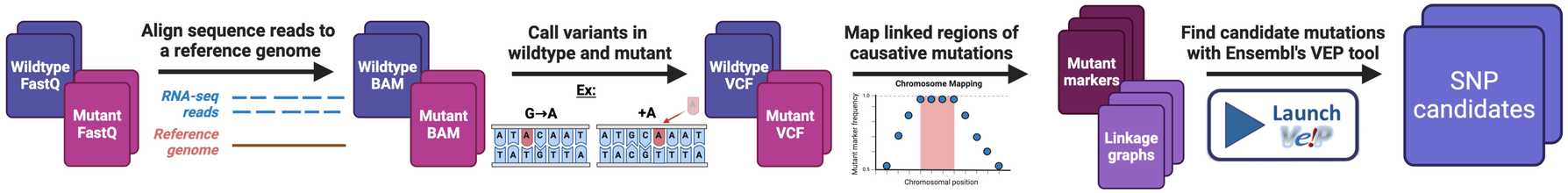



**Figure 1: Bioinformatic workflow for mapping pipeline.** This pipeline uses publicly available tools for alignment, variant calling, variant effect prediction, and differential expression analysis. Custom-built R scripts output maps of linked regions and lists of candidate genes based on the predicted effects of each variant.

## 3. Identification of linked regions in control and novel datasets

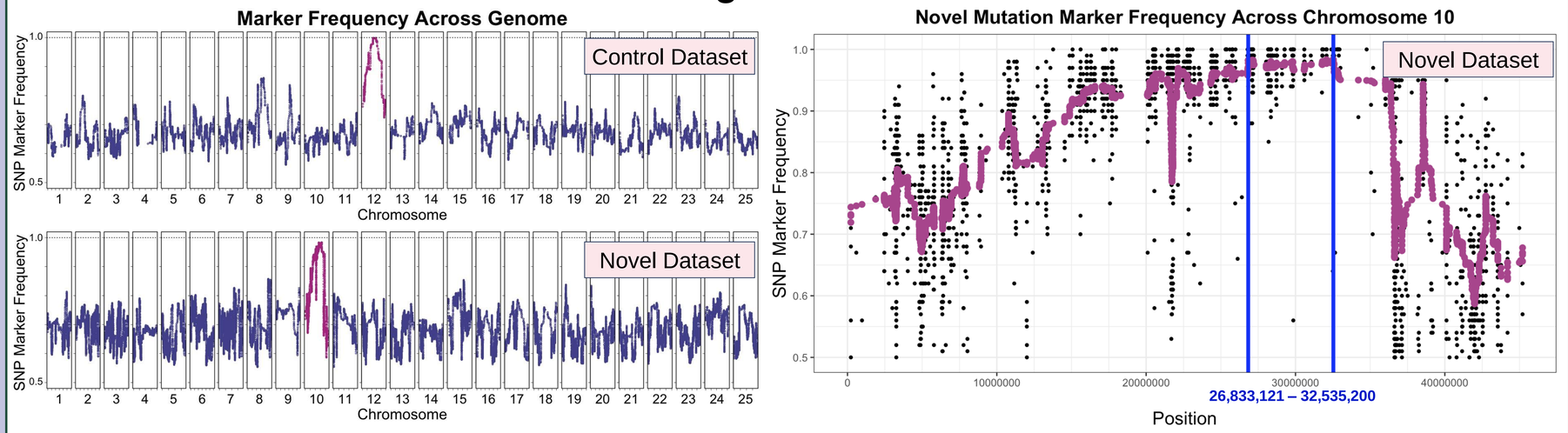



**Figure 2: Linked mutations in control *hoxb1b* and novel mutant dataset.** Purple and magenta markers represent the sliding window average of the SNP's frequency and black markers represent all SNPs across the chromosome. Linked mutant region located on Chromosome 10 is denoted by the vertical blue lines.

## 4. Identification of candidate mutations

**Table 1: Top Five SNP & INDEL Candidates from Chromosome 10**

| Gene | Position | Count | Freq. | Consequence | Function |
|---|---|---|---|---|---|
| *cadm2a* | 23213868 | 123 | 0.73 | intron variant | cell adhesion |
| *cltca* | 28335761 | 40 | 0.88 | intron variant | protein transport |
| *nsd3* | 20586604 | 29 | 0.93 | stop gained | histone methylation |
| *pcdh1g29* | 21799222 | 22 | 0.59 | stop gained | cell adhesion |
| *frem2a* | 25986164 | 18 | 0.83 | stop gained | cell adhesion |

## 5. Improvements to tool performance

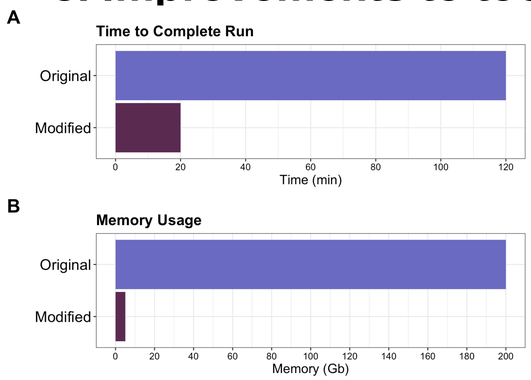



**Figure 3: Performance comparison of modified tool to original tool.** (A) Time to map genome before and after custom modifications. (B) Memory improvement for mapping genome before and after custom modifications.

## 6. Future Directions

- Experimentally test candidate genes to determine which is causing the observed phenotype
- Build and launch an accessible web-based application for non-bioinformaticians
- Apply pipeline to other model organisms

## 7. Acknowledgements

We would like to thank the UO GC3F, our mentors and peers within the Knight Campus Graduate Internship Program for their unwavering support, and the NIH (R21NS135433) for project funding. This work benefited from access to the University of Oregon high performance computing cluster, Talapas.

1. A. C. Miller, N. D. Obholzer, A. N. Shah, S. G. Megason, C. B. Moens, RNA-seq-based mapping and candidate identification of mutations from forward genetic screens. *Genome Res* **23**, 679–686 (2013).
2. M. E. Bowen, K. Henke, K. R. Siegfried, M. L. Warman, M. P. Harris, Efficient mapping and cloning of mutations in zebrafish by low-coverage Whole-Genome Sequencing. *Genetics* **190**, 1017–1024 (2012).