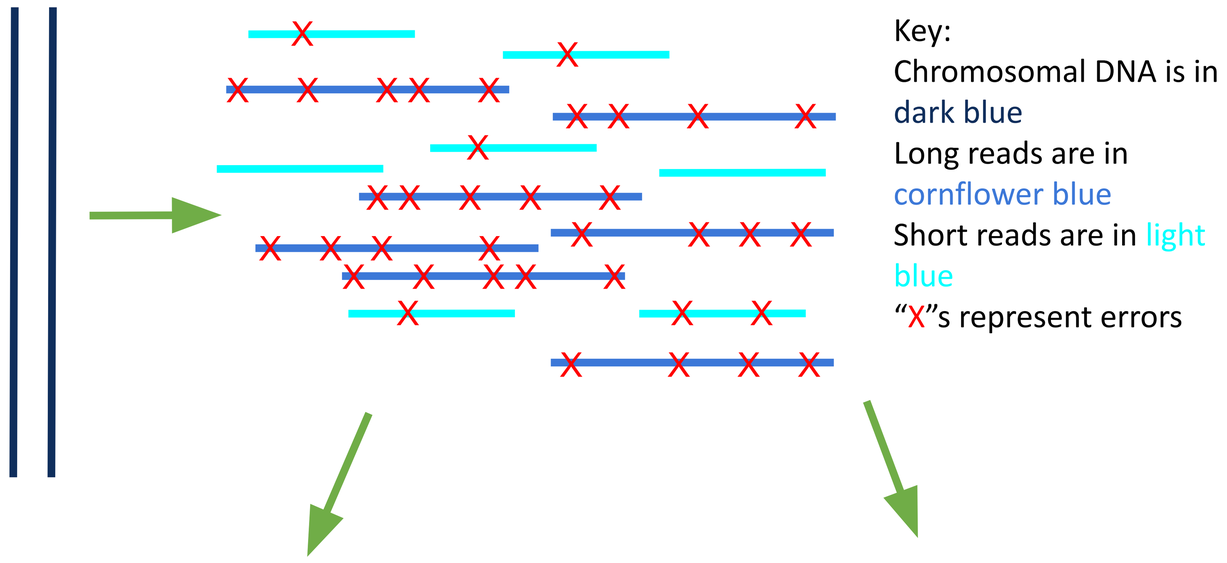
A Computational Analysis of Hybrid Genome Assemblies.

Joseph Walewski[1,2], Stephen Douglass[2,3]
Biochemistry, Cellular, and Molecular Biology[1]; Computer Science[2]; Biology[3]

Introduction.

As the central dogma of molecular biology states DNA is transcribed to RNA and then translated to protein. Due to this the genome of an organism yields key insights into the molecular machinery it operates, from regulatory sequences to alternate splicing sites. Additionally, the genome can yield unique insights into phylogeny. Current technology supports two different ways of obtaining genomic data: either with short but accurate reads, or with long and inaccurate reads. Hybrid genome assembly attempts to bridge this gap by combining the accuracy of short reads with the ability of long reads to span repetitive regions, a critical feat as genome size gets larger. Although the cost of genomic sequencing has fallen approximately 10,000 fold over the past twenty years, many aquatic taxa, including lungfish and salamanders, have such large genomes that *de novo* genome sequencing efforts still cost tens of thousands of dollars. **Therefore, our group has been interested in identifying the most cost efficient way in assembling a large eukaryotic genome.**

Core problem: **How to best combine short but accurate and long but inaccurate reads to generate the best assembly possible?**



Preliminary Results.

Figure 1. The fraction of the original genome recovered compared to the error rate for long read correction (FMLRC2) assemblies. Three key trends exist - first, **the fraction of the original genome recovered is strongly positively correlated with the long read coverage used**. Next, **the short read coverage used is strongly negatively correlated with the error rate and weakly positively correlated with the fraction of the original genome recovered**. Third, despite the genome size of *D. rerio* being approximately ten times larger than *A. thaliana*, **there is no significant difference in the fraction of original genome recovered when using FMLRC2 and the same short and long read coverage levels.** Preliminary data from *H. sapiens* suggests this continues up to a genome size of at least 3.2 Gb.

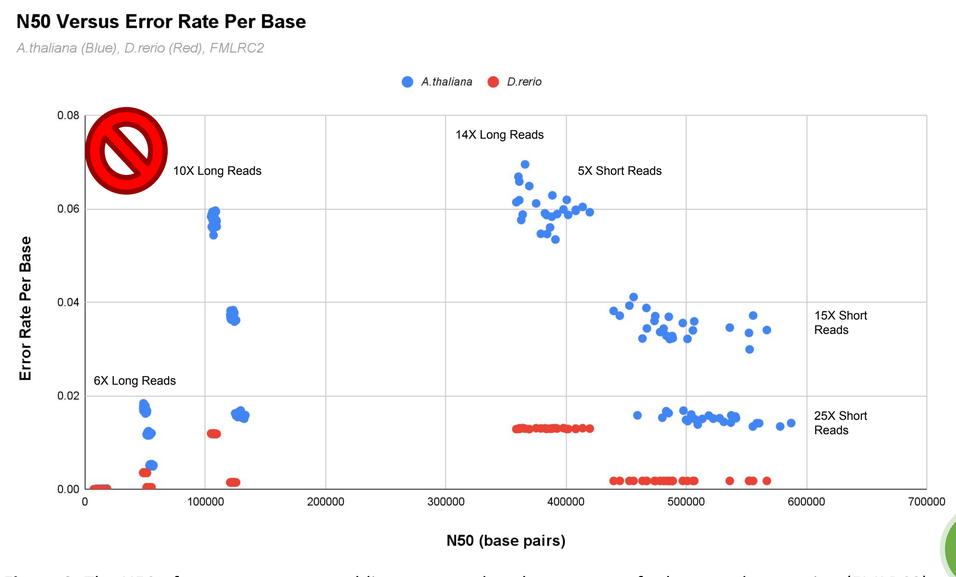


Figure 2. The N50 of new genome assemblies compared to the error rate for long read correction (FMLRC2) assemblies. Three key trends exist - first, **the N50 of the assemblies is very strongly correlated with the long read coverage used**. Next, **the short read coverage used is strongly correlated with the error rate and weakly correlated with the fraction of the original genome recovered**. Third, despite the genome size of *D. rerio* being approximately ten times larger than *A. thaliana*, **there is no significant difference in N50 when using FMLRC2 and the same short and long read coverage levels.** Again, preliminary data from *H. sapiens* suggests this continues up to a genome size of at least 3.2 Gb.

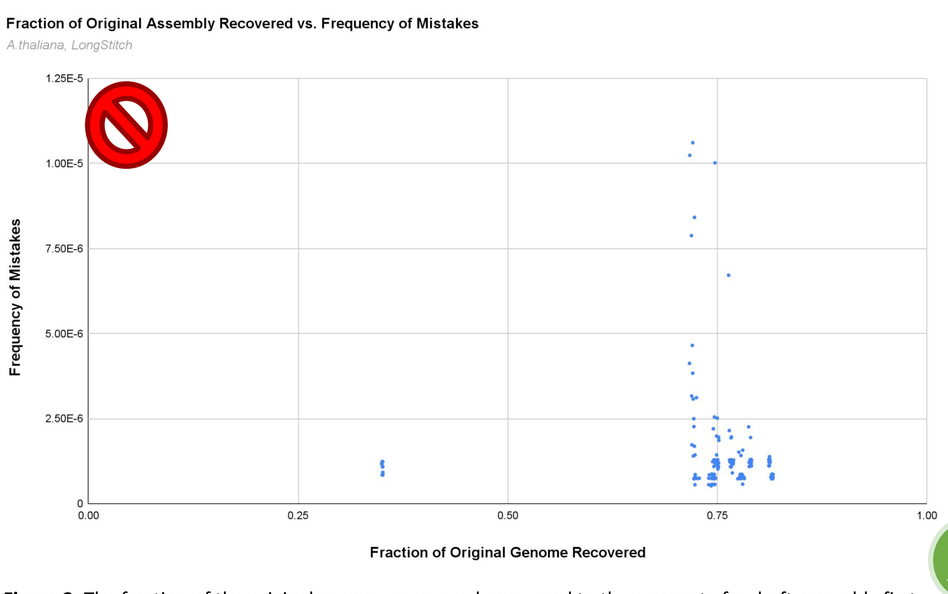


Figure 3. The fraction of the original genome recovered compared to the error rate for draft assembly first (LongStitch) assemblies. There was very little correlation observed, with the vast majority of assemblies returning 74 and 80% of the original genome. Due to this, our priority was investigating FMLRC2 further, so less complete results currently exist for LongStitch.

- Option 1 (FMLRC2):

FMLRC2 loads long reads into memory and then maps short reads onto them using two implicit deBruijn graphs with different kmer lengths (decided dynamically). After correcting the reads minimap2 generates pairwise alignments, which miniasm then uses to assemble the genome.

- Option 2 (LongStitch):

LongStitch generates a draft assembly with BCALM2, a short read - only assembler, and then uses Tigmint, NTlinks, and Arcs-long (optionally, we opted to use it) to map long reads to the draft assembly and identify gaps in the short read assembly that could be "stitched" with the available long reads.

- We asked this question in four organisms - *A.thaliana, D.rerio, H.sapiens, & P.waltl* (The iberian ribbed newt, of interest to us as we wish to sequence a salamander genome). Previously published genomes and sequence reads for each species were downloaded and RESEQ (short reads) and PBSIM2 (long reads) were used to simulate varying levels of coverage for each read type and species. To ensure reproducibility of results, 25 replicates of each assembly were generated. Completion and accuracy were assessed with a combination of GSAlign and an in house program.

Conclusions and Future Directions.

FMLRC2 can produce significantly more complete assemblies. At present, data for all replicates only exists for *A. thaliana,* and the majority of the final data exists for *D. rerio*. Our *in silico* analysis demonstrates that FMLRC2 generally outperforms LongStitch in both species (the only exception is when Pacbio read coverage is very low). Given these results, as we transition to our own sequencing of the eastern newt genome it is likely FMLRC2 will be the assembler of choice and therefore we will request coverage amounts optimized for that assembler.

The second phase of Professor Douglass' and I's research will investigate the newly completed eastern newt genome by identify homologues of genes already known in other species. The amount of similarity we see here, along with other key sections of the genome such as repetitive regions, transposons, and telomeres, will allow us to determine the synteny (pattern of chromosomal conservation between species) and therefore the phylogeny of the species. Since sequencing the eastern newt genome will finally allow for comparative genomics in the newt family, we may uncover some interesting results that could potentially change our view of how various newt species are related.

Critically, this could change our perception of how similar wound healing is between various newt species. Regardless of the result, it will have wide ranging impacts: either newt regeneration is highly conserved, meaning that it is more ancestral than previously anticipated; or it varies significantly between eastern and Iberian newts - which would suggest the existence of multiple pathways for extreme regenerative potential in newts. One day this may indicate flexibility in therapeutic options.

Acknowledgements.

I want to thank Professor Douglass for his extreme commitment to what was originally a proposal to extend a BIO 120D final project. This work has grown so far beyond that, to the most educational experience of my undergraduate career. Additionally, I would like to thank Connecticut College's Summer Science Research Program, for offering stipend money, without which this project would not have been possible (the stipend was used to build a research computer which was absolutely essential for the completion of this project). I would also like to thank the Institutional Animal Care and Use Committee for their guidance and support. I would also like to thank our partners beyond the Connecticut College Community the Mill River Committee, the Connecticut Audubon Society, the Town of Fairfield Department of Conservation, and Connecticut's Department of Energy and Environmental Protection (DEEP) for legally approving this research.