Improving the Draft Assembly of the Horse Genome

Megan Smedinghoff, smeds@umd.edu

Advisor: James A. Yorke, yorke@umd.edu

October 11, 2007

Abstract

I aim to improve the draft genome of the horse that was produced by Broad Institute in early 2007. My strategy is to begin by producing a second assembly of the horse using the Celera Assembler. After generating a new assembly, I intend to use existing University of Maryland software to reconcile the two assemblies and produce a third that is more accurate than either of the two drafts.

Background

In February 2007, the Broad Institute released a draft assembly of the horse genome (*Equus Caballus*). The release was the culmination of a $15 million project funded by the National Institute of Human Genome Research and the National Institute of Health. The draft genome will allow the equine research community to better understand diseases that affect horses. Additionally, the release of the horse genome has caused some excitement in the human genome research community. There are 80 known conditions in horses that are analogous to disorders in humans (including arthritis and allergies). Many think that analyzing the horse genome will give insight into treating these conditions in humans. Finally, Broad Institute also released a database of SNPs (single nucleotide polymorphisms) that can be used to identify different breeds of horses. Researchers expect this database to help locate links between genomic code and physical characteristics.

Project Goal

Broad Institute assembled the draft using the Arachne assembler. I propose to reassemble the horse genome using the Celera assembler. The two assemblies will differ since the two assemblers rely on different algorithms (the main difference being that Arachne uses mate pair information at the very beginning of the assembly process whereas Celera uses it during contig assembly). After producing the Celera assembly, I propose to reconcile the two assemblies using existing University of Maryland software. The hybrid assembly will be more accurate than either of the individual assemblies.

Implementation

I intend to use the genome cluster at University of Maryland to implement my proposal. Much of the project will involve using existing software. In addition to the Celera assembler, I also intend to use existing University of Maryland software (vector trimmer, overlapper, and reconciliation programs). While much of software already exists, the assembly of a large mammal is not trivial and will not doubt require me to write additional programs. I will use Perl to write these smaller scripts.

Verification

An important part of any programming project is to be able to verify the results. Verification if particularly tricky in the case of genome assembly, though. The repetitive nature of genomes and the uncertainty of sequencing techniques guarantee that the genome will not be completely correct. There are resequencing methods that give very accurate results, but these methods are prohibitively expensive in most cases. In my case, I will compare my Celera assembly to the Broad assembly and make sure they are similar. I will use Mummer to do the comparison between the two genomes. I expect there to be about 1.5% difference between the two assemblies.

Secondary Goals

My primary goal is to produce an improved assembly of the horse. In addition to producing an assembly, I would also like to use some of the software that has been created by The University of Maryland in the past few years. There are several programs that either have not been used yet or have not been tested on large genomes. I am interested in running some of these programs to see how they affect my assembly. Hopefully I will be able to pass on some useful comment to the authors regarding the performance of these programs.

Timeline

Much of this project involves using software that I have never worked with before. Consequently, I am unsure how long it will take to perform certain tasks. In many cases, I am not yet sure what tasks need to be performed in order to run certain pieces of software. However, from discussing my project with my advisor, I feel that the following major goals are realistic:

Fall 2007: Produce a Celera Assembly

January 2008: Compare the Celera and the Arachne Assemblies

Spring 2008: Produce a reconciled assembly and write up my results

Conclusion

The draft sequence of the horse was released in early 2007. The release provides many opportunities for the genome research community. Improving the draft will benefit equine researches who are interested in studying diseases that affect horses. Improvement will also benefit those who are interested in studying human diseases and conditions that are also present in horses. Finally, improving the assembly will help test the many programs that are used in genome assembly. I intend to create an improved assembly of the horse by producing a draft with the Celera assembler and then reconciling the results with Broad's draft assembly. I hope to release the improved genome by the end of May, 2008.

References

NIH News (February 7, 2007) Horse Genome Assembled.

http://www.nih.gov/news/pr/feb2007/nhgri-07.htm

Broad Institute Horse Genome Project Webpage

http://www.broad.mit.edu/mammals/horse/