

A Genome Sequence Resource for Barley

Matthew Alpert¹, Steve Wanamaker², Denisa Duma¹, Raymond D. Fenton², Yaqin Ma²,
Gary J. Muehlbauer³, Stefano Lonardi¹, Timothy J. Close²

¹Department of Computer Sciences, University of California, Riverside, CA, 92521

²Department of Botany & Plant Sciences, University of California, Riverside, CA, 92521

³Department of Agronomy & Plant Genetics, University of Minnesota, St. Paul, MN, 55108

A new assembly of genome sequences of Morex barley (version 0.05) has been posted on the internet for public access. This is one of several new information resources under development in a project entitled, “Advancing the Barley Genome”, supported by the Agriculture and Food Research Initiative of USDA’s National Institute of Food and Agriculture. Sequences may be queried using the BLAST interface at www.harvest-blast.org and retrieved from <http://harvest-web.org/utimenu.wc> following the instructions for “Get a FASTA file of Selected Barley Genome 0.05 Sequences”. Alternatively, the entire FASTA file containing 1.19 Gb of assembled genome sequences is available by sending a request to timothy.close@ucr.edu and agreeing to avoid collisions with publication plans of the project members. Alternative versions of the assembly after masking for repetitive sequences also are available on request.

Morex barley genome sequences were assembled using SOAPdenovo by 4th year UC Riverside (UCR) undergraduate student Matt Alpert with assistance in data cleaning by programmer Steve Wanamaker and analyses of the assembly by other UCR team members, Duma, Close, Wanamaker and Lonardi. Matt Alpert is presently thinking about his future educational and career options after he will graduate from UCR in June 2012. The sequences were generated using Illumina library methods and both GAII and HiSeq instruments at three locations including Ambry Genetics (Aliso Viejo, California), University of Minnesota (Muehlbauer location) and the UCR Institute of Integrative Genome Biology. Five independent whole-genome standard-size libraries (~300-350 bp) and three long insert libraries (2, 3 and 5 kb) were prepared by either Yaqin Ma at UCR, others at U Minnesota, or Ambry Genetics. Morex nuclear DNA was isolated by Amplicon Express (Pullman, Washington) from tissue generated by Raymond Fenton from a stock of 100% homozygous Morex barley seeds. After trimming off adapter and low quality sequences, an input dataset of 164 Gb of barley genome sequence, about 30x coverage, was used for assembly. A k-mer frequency analysis indicated that the depth of coverage of the assembled sequences was 24x. Highly repetitive DNA generally does not assemble well, so the resulting assembled sequences include only about 22% of the barley genome. However, the assembled sequences in “Barley Genome 0.05” include more than 90% of all previously identified barley genes and are useful for a number of purposes including PCR primer design and identification of introns and gene regulatory sequences adjacent to expressed genes.

This newly updated partial genome sequence of barley is just one outcome of several rapidly developing projects involving members of the International Barley Sequencing Consortium (IBSC). News items about the barley genome, including publications and pre-publication access to new genome resources, are frequently posted on the IBSC website www.barley-genome.org to provide barley researchers with timely access to information intended to help meet various objectives.

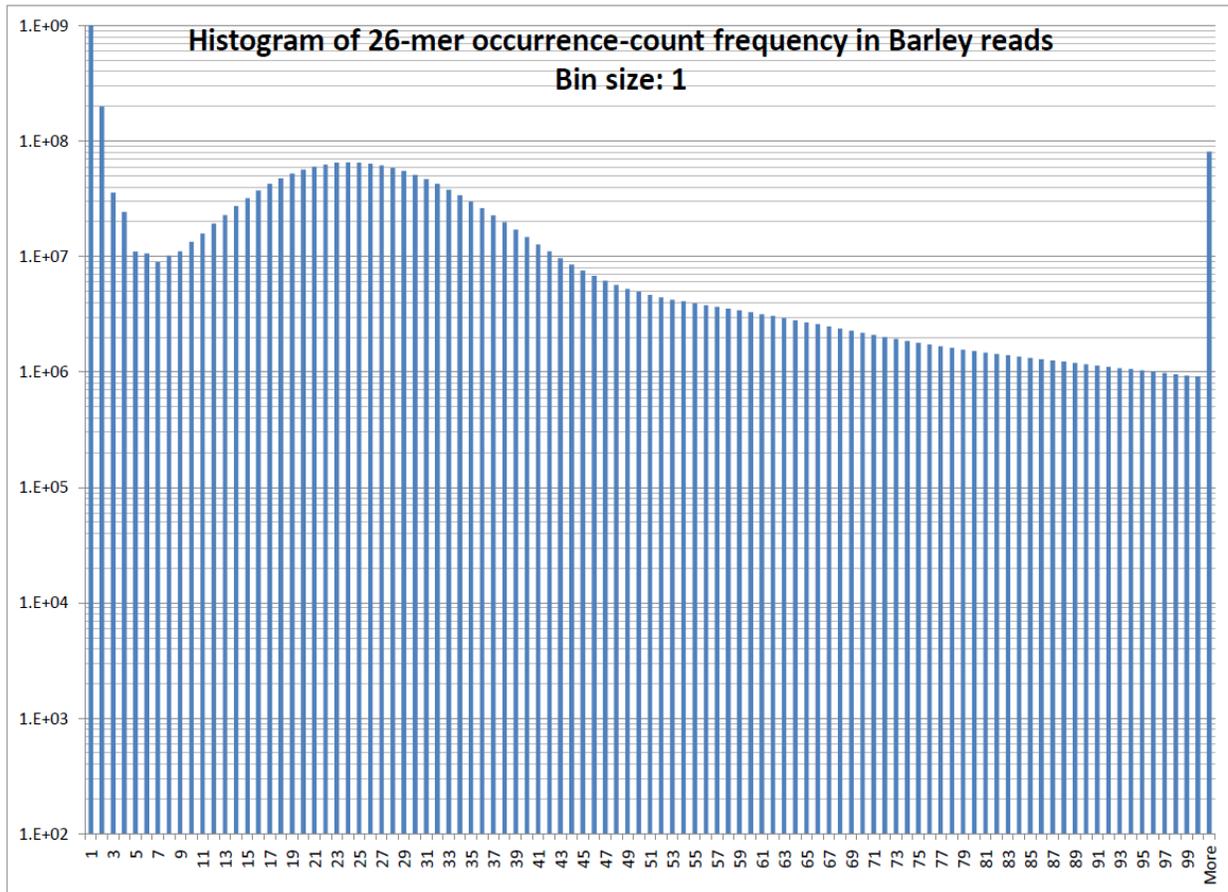


Figure 1. Depth of coverage. Each quality-trimmed and adaptor-trimmed read from the sequencer was used to generate a list of 26-mers, each offset by one base. For example, a 91-base sequence would yield 66 26-mers. This is referred to as "hashing" the sequence information. The resulting 26-mers were used to create a comprehensive "hash table" for the entire set of reads, with the frequency of occurrence of each 26-mer also determined. The histogram displays the number of 26-mers (y axis) in each frequency group (x axis). The peak in the range of 21-27 (mode 24) provides a measure of the actual depth of nuclear genome coverage of the assembly because it represents the depth of coverage of each unique sequence in the genome. A length of 26 bases was sufficient for this analysis because the odds of any particular 26-mer occurring by chance is only 4 to the 26th power, or 1 per 4.5×10^{15} 26-mers, whereas the input dataset was composed of only 3.1×10^9 26-mers.