**Genomic Assembly, Usage, Problems and Limitations.**

Author: Grant Chrapla
Major: Microbiology/Cell and Molecular Biology
Department of Microbiology and Molecular Genetics, Oklahoma State University, Stillwater, OK 74078, USA

**Key Words:**

Genomics, gene assembly, bioinformatics, machine learning

**Genetic knowledge is leading a frontier of research towards designing custom solutions for problems we might experience during our lives. An organism’s genome is akin to a book, it is a publication that is uniquely them and defines what they are. Regarding medicine, it can be assumed that common ailments have vastly complex interactions in how they are expressed in an individual. Rare genetic disorders are usually defined with singular genetic events. These genetic events can be as simple as a base mutation to as complex as large deletions, duplications and inversions. After our discovery of DNA being the chemical component of inheritance in reproduction around the 1950s, an international effort known as the Human Genome Sequencing Consortium was established in the 1990s to give researchers the first draft of a human genome. A finished genome (book) by itself means nothing to us. Alignment of nucleic acid intermediaries such as cDNAs, mRNAs, tRNAs and sRNAs to draft genomes gives us a gene (word) count. Manipulation of these genes could lead to observable functions (definitions). Advanced studies try to associate gene to gene interactions in metabolic-pathways (sentences). Developmental stages in organisms can observe differences in DNA methylation patterns (paragraphs) and histone packaging (chapters). Our comparative studies to better understand genomes now requires increasingly large amounts of data. Hundreds of these example datasets could be needed to increase confidence in answers we find for the questions that are asked. Techniques to produce and process quality data needs to be refined further to reduce the time needed to research questions.**

**Introduction**

One of the largest and most complex reference genomes completed to this date would have to be bread wheat. The International Wheat Genome Sequencing Consortium published this draft late 2018. This genome is over 4,000x the size of the human genome and was completed in a similar time frame of about 14 years. This project started during a drastic change in sequencing technologies, commonly referred to as next generation sequencing (NGS). This assembly was derived from carefully curated BAC libraries (approx. 2,000 kbps) of separated sub genome (A, B and D) chromosomes fractured into over a quarter million BACs to cover the entire genome. Most of the sequencing was completed on Illumina’s bridge amplification technology, their sequencers can produce reads of about 300 bps. Shot-gun sequencing approaches, of sufficient read depth, were assembled into contigs that create increasing larger consensus sequences.

Marketability of data drives research and gets inquiries funded. Wheat is one of the world’s commodity crops that could greatly benefit from increased drought tolerance, disease resistance, increased nutritional content and crop yields. Along with feeding the world it is also used in the industry of popular fermented malt beverages. About commodities, a countries’ population is also a market that drives research. One large killer is cancerous health disorders. Cancer mostly knows no bounds and can afflict populations in every socioeconomic level making it an active area of research. The single unifying market that drives this research is the computational technology that gives us the tools to decipher what could be centuries of data down to a few hours a day to years at a time. Developing new tools that scale to assemble genomes of all sizes could greatly reduce the time needed for smaller genomes. One research group said they took 200,000 compute hours (8,334 days or 23 years for one core cpu) to complete one wheat genome draft. If that same workload was optimized to run on a 128 core cpu compute cluster, it could take about 65 days. Assembling a human genome in a similar fashion could take 24 minutes.

**Discussion**

Genomics is only a small subset of Omics related studies. There is a movement to integrate all types of omic related research to further understand life. The ability to utilize data is far from perfected and a complete reimagining of data processing is needed. I feel as of 2019, referencing NGS as a blanket term for sequencing technologies is a disservice. Single reads, paired reads, optical florescence, digital voltage potential, short reads, long reads and scalability gives real meaning to sequencing. This better characterizes data that is generated and assists with leading a new generation in how to manage and utilize such technologies. To give an example of data characterization, Oxford Nanotech’s Minion sequencing technology works well for this. The flow cells of the Minion recognize a voltage potential through a pore. The output is visualized as a noisy line that rises and falls as a DNA fragment travels through the pore. Depending on this flow cell revision, this could be a line based off all nucleotide combinations 5-8 nucleotides in length. The first voltage read should define what the next potential base call is over time. An artificial neural network (ANN) implementation on this base calling could give many candidate sequences. As the fragment progresses, there is a chance, more distinct readings could differentiate one candidate from another until one sequence is identified as valid. The algorithm then can score each reading and predict the next potential candidate for base calls. If the reading differs greatly from the prediction, environmental factors could a produced an erroneous result, if the erroneous reading did not progress longer then the set threshold in the 5-8 base calls then the error could be overwritten with the corrected values. As the ANN experiences issues with the reads, and independent human assessment found no problems, it learns on how to judge further experiences. So, it then becomes important to be able to produce quality samples for trained neural networks. If you ever experienced great variability in DNA yields from samples given the same conditions under the same procedure, you’ll need to ask yourself what was the cause? Reproducibility of data is needed to give confidence in the work produced in research.

**References**

Zuzana Tulpová-Ming-Cheng Luo-Helena Toegelová-Paul Visendi-Satomi Hayashi-Petr Vojta-Etienne Paux-Andrzej Kilian-Michaël Abrouk-Jan Bartoš-Marián Hajdúch-Jacqueline Batley-David Edwards-Jaroslav Doležel-Hana Šimková “Integrated physical map of bread wheat chromosome arm 7DS to facilitate gene cloning and comparative studies”. New Biotechnology (2019)