

A New Method for Long Range Scaffolding using Optical Mapping

Nianqing Xiao, Ryan N. Ptashkin, Thomas S. Anantharaman, Bin Zhu, Deacon Sweeney, John K. Henkhaus
OpGen, Inc., 708 Quince Orchard Rd, Gaithersburg, MD 20878

Introduction:

Rapid improvement in throughput and accuracy of sequencing technology has provided the opportunity to sequence the complete genomes of thousands of organisms, and therefore the potential to gain great biological and clinical insight. However, due to the limitation of the sequence read length, it still remains challenging to correctly resolve the genomic regions of complex nature. Whole genome sequencing project often produce hundreds even thousands of contigs and scaffolds. Although the sequences typically cover a majority of an organism's genome, but the relative order and orientation is difficult to determine. Assembly of large genomes, such as plant and animal genomes, is further burdened by presence of repetitive regions and other complex genomic structures.

Various technologies and techniques that provide long range genomic information have been developed to help overcome the problem. So far, these approaches tend to be labor intensive and time consuming, or require prior knowledge about the genome that is not easily available.

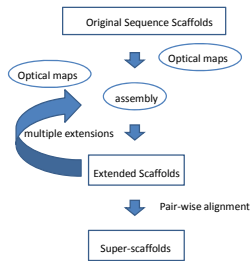
Optical mapping technology generates restriction maps from single DNA molecules *de novo*, typically in the range of 200 kb up to 1 Mb. It has been successfully utilized to compare structural difference between microbial genomes, detect errors in sequence assemblies, and determine the order and orientation of sequence scaffolds. Recent improvements in Optical mapping technology have made it feasible to generate large amounts of high quality Optical Mapping data in an automated and consistent manner using Argus™ Optical Mapping System. Moreover, a computational framework that bridges and orientates sequence scaffolds with optical maps has been developed, and a software pipeline based on this framework has been implemented. Thus, application of Optical Mapping in finishing of large complex genomes such as plant and animal genomes become practical.

In this study, we use human genome as model system to test the ability of our approach to join large sequence scaffolds that are up to a few hundred kilo-bases apart from each other.

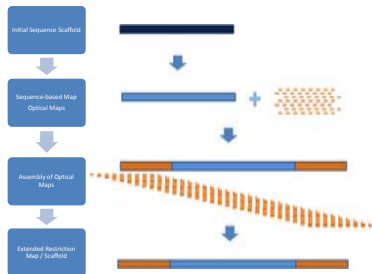
Computational Methods:

The basic approach was to use Optical Maps to extend sequence scaffolds, create overlapping regions between adjacent scaffolds and join them based on alignments between the extended scaffolds. Briefly, the sequence scaffolds were first converted into restriction maps by *in silico* restriction enzyme digestion. The resulting *in silico* maps were used as seeds to identify single molecule restriction maps of DNA from the corresponding genomic regions. These single molecule maps were subsequently assembled together with the *in silico* maps, producing elongated consensus maps (extended scaffolds). The low coverage regions towards both ends of the extended scaffolds were trimmed off to maintain high extension quality. To generate longer extensions, the alignment-assembly process was typically iterated up to 4 times, using the extended scaffolds as seeds for the next iteration. All of the extended scaffolds were then aligned to each other. Any pair-wise alignments above a certain confidence threshold were considered as initial candidates for scaffold connection. Heuristics were used to filter out spurious alignments, and conflicts among the remaining alignments were finally resolved based on the significance of alignments.

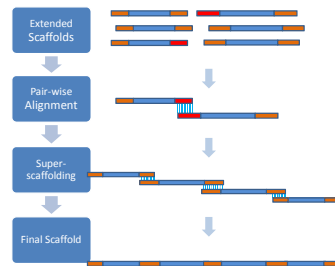
Overall Computational Workflow



Step I: Scaffold Extension using Optical Maps



Step II: Connecting Extended Scaffolds

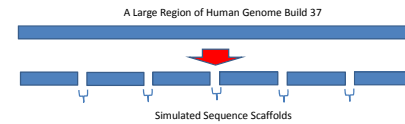


Data Generation:

Optical Mapping data from human blood cell DNA have been collected from 58 MapCards using Argus Optical Mapping system. The average time for image collection is 1.5 hours per card, so the total system time for data collection is 90 hours. The DNA molecules were marked up and restriction fragment size was determined by image processing in parallel with data collection. The total size of single molecule restriction maps (SMRMs) (> 250 kb) is about 580 Gb, averaging about 10 Gb per MapCard. About 10% of SMRMs aligns to the reference genome (HG Build37) with high stringency, averaging about 1 GB per MapCard.

Simulated Test Sequence Scaffolds:

To test the ability of our approach to join large sequence scaffolds, simulated test sets were generated from human genome sequence build 37 by randomly splitting uninterrupted sequences into large sequence scaffolds. Specifically, 15 randomly selected genomic regions from 12 chromosomes were into 6 artificial scaffolds each. A portion of sequence of a pre-specified size was removed between two scaffolds to mimic "gaps" between scaffolds (as illustrated below). Each test set has 90 scaffolds with artificial 75 gaps. Six test data sets with gap sizes of 10, 20, 50, 100 and 200 kb, respectively, were generated for testing.

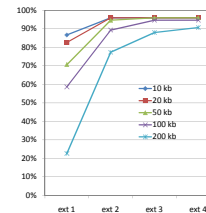


Results:

The randomly generated sequence scaffolds were combined with the Optical maps generated using Argus system, and joined using OpGen large genome bioinformatics pipeline. The relative location and orientation between the joined scaffolds were compared to the "truth" that was inferred from original reference sequence.

As shown below (left panel), multiple extensions were required to generate sufficient overlapping between the scaffolds and achieve maximum gap closure rate, especially for larger gaps (e.g. 100 kb and 200 kb). At the end of fourth iteration (right panel), over 90% of gaps of 200 kb have been closed, while the closure rate of smaller gaps reached 96%. No incorrect joining between the scaffolds has been observed at any iteration of any gap size tested.

Gap Closure Rate vs. Extensions



Final Gap Closure Rate and Accuracy

Gap Size	Closure Rate	Error Rate
10 kb	96%	0
20 kb	96%	0
50 kb	96%	0
100 kb	95%	0
200 kb	91%	0

Conclusions

In this study, we have demonstrated that:

1. Optical Mapping technology can be used to effectively and accurately bridge sequence scaffolds that are up to 200 kb apart in large complex genomes.
2. The Argus Optical Mapping System is capable of rapidly generating a large amount of high quality Optical Mapping data to support large genome finishing.
3. The OpGen computational pipeline provides an efficient solution for long range super-scaffolding using Optical Mapping data.
4. Optical Mapping can be coupled with next gen sequencing technologies to significantly reduce the complexity associated with finishing large genomes and improve the quality of large genome sequences.