

Application of Argus® Whole Genome Mapping System in Analysis of Large Complex Genomes

Nianqing Xiao¹, Bin Zhu¹, Thomas S. Anantharaman¹, Deacon Sweeney¹, Yunhu Wan¹, Ryan N. Ptashkin¹, Xun Xu², John K. Henkhaus¹.

1) OpGen, Inc., Gaithersburg, MD; 2) Beijing Genome Institute, Shenzhen, China.

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 a challenge 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.

Whole Genome Mapping 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 Whole Genome Mapping technology have made it feasible to generate large amounts of high quality single molecule maps in an automated and consistent manner using Argus™ Whole Genome Mapping System.

In this report, we present our recent works in application of Whole Genome Mapping in long-range scaffolding of human and animal genomes, and detection of structural variations in human genomes.

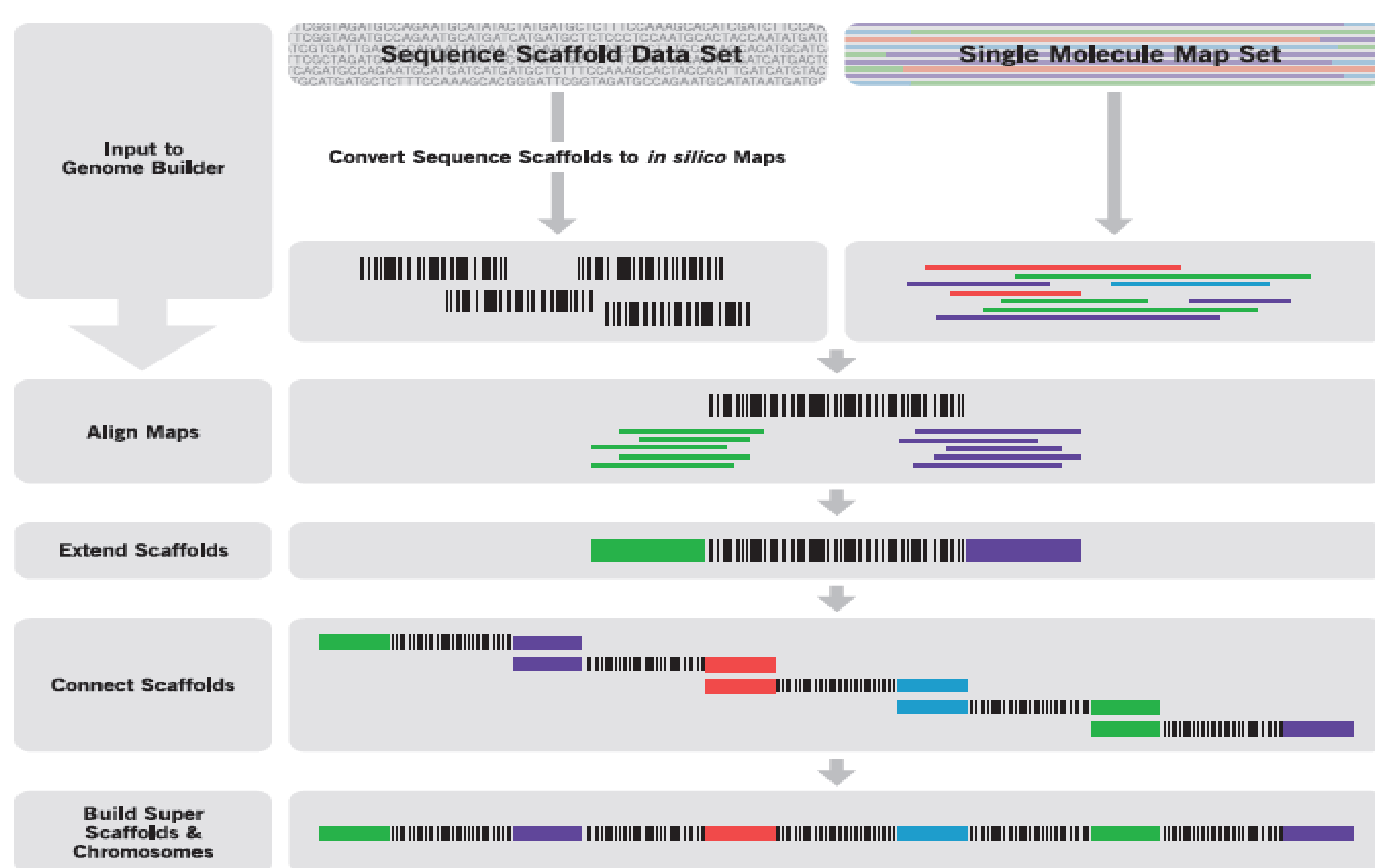
Part I. Long-range Scaffolding with Genome Builder

We have developed a computational framework for long-range scaffolding of large complex genomes, and implemented it in the Genome Builder software package. To test the utility of Whole Genome Mapping in sequencing large genomes, we have conducted proof-of concept studies with the human and goat genomes.

Computational Methods:

The basic approach was to use single molecule 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.

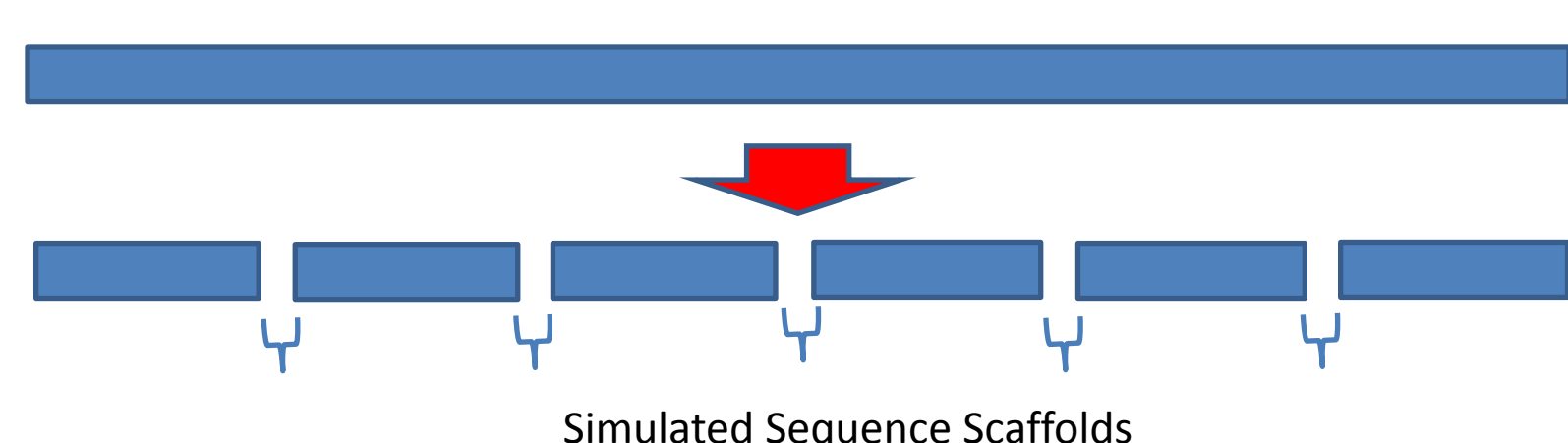
Genome Builder Computational Workflow



Simulated Study on Human Genome

Simulated Sequence Scaffolds:

To test the performance of our approach in large genome sequence assembly, simulated test data sets were generated from GRCh37 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.



Data Generation:

Single molecule maps have been collected from human blood cell DNA with 58 MapCards using Argus Whole Genome 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 (GRCh37) with high stringency, averaging about 1 GB per MapCard.

Scaffolding Efficiency and Accuracy:

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, 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 reaches 96%. No incorrect joining between the scaffolds has been observed at any iteration of any gap size tested.

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

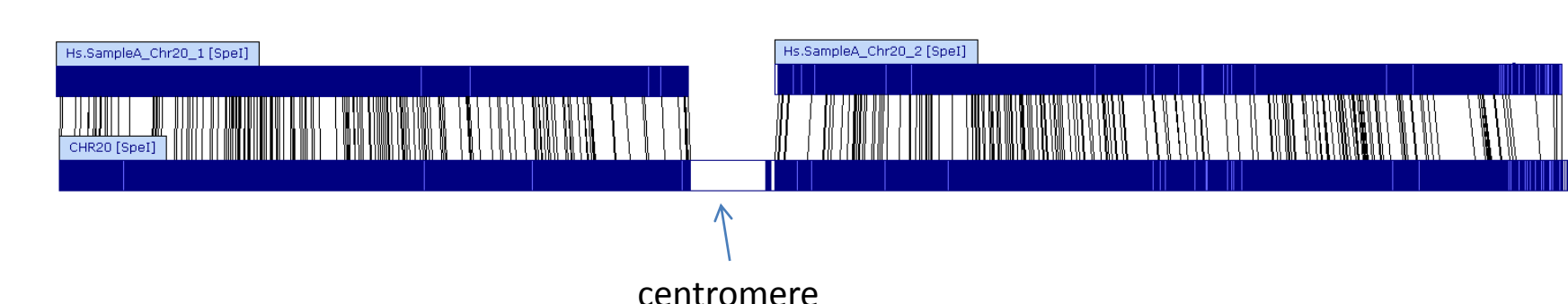
Goat Genome Sequencing Project

This approach was applied to goat genome *de novo* sequencing project at BGI where whole genome, shotgun sequence information was obtained using the Illumina platform. Whole Genome Mapping was independently performed at OpGen from frozen, epithelial cells. As shown below, scaffold N50 was improved nearly 8-fold (from 2.2 Mb to 16.9 Mb). The N90 was similarly improved from 0.5 Mb to 2.8 Mb, reducing the total number of scaffolds that cover 90% of the genome from 1236 to 181.

	BGI Input	Genome-Builder Output
N50 (MB)	2.29	16.89
N80 (MB)	0.91	6.23
N90 (MB)	0.52	2.83
Scaffolds (90% genome coverage)	1236	181
Scaffolds (from scaffolds >200kb)	N/A	1284

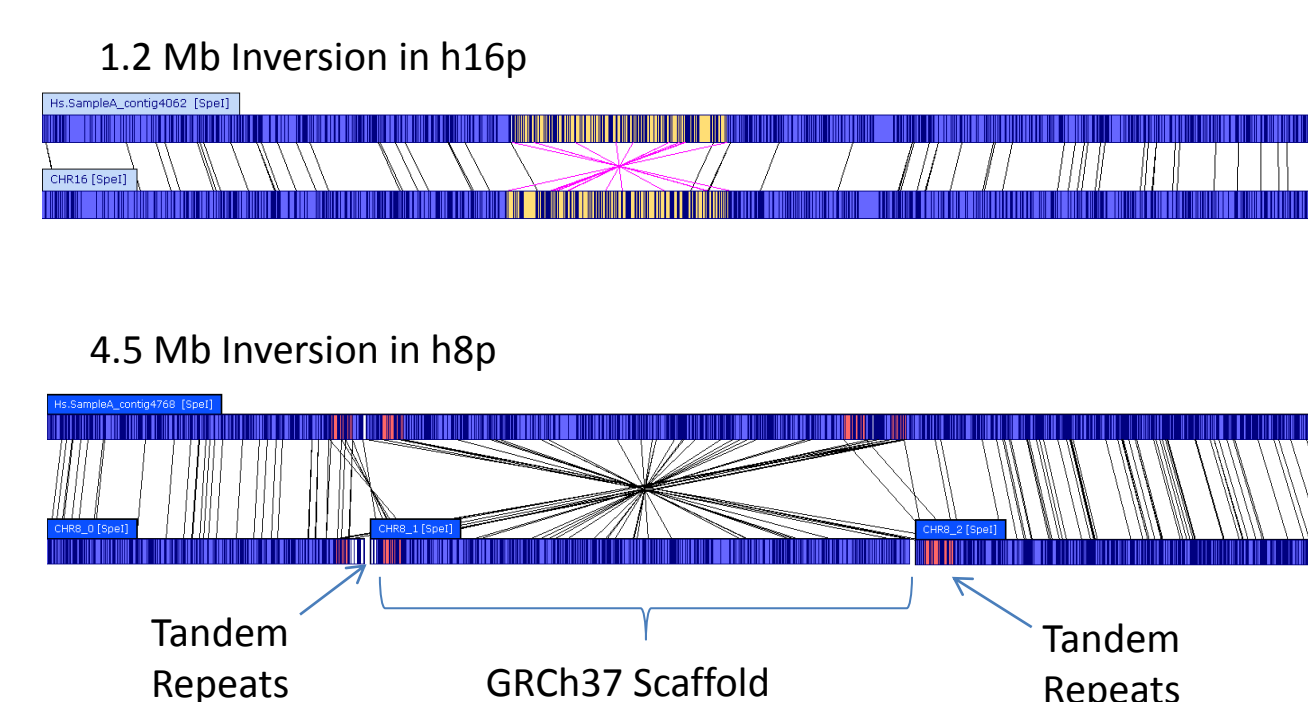
Part II. Structural Variation Detection

Single molecule restriction maps from a human DNA sample are assembled into consensus maps using a new computational pipeline developed at OpGen. The resulted consensus maps span large genomic regions, up to individual arms of chromosomes (e.g. chromosome 22 as shown below), which provides global, yet information-rich view of genomic structure.

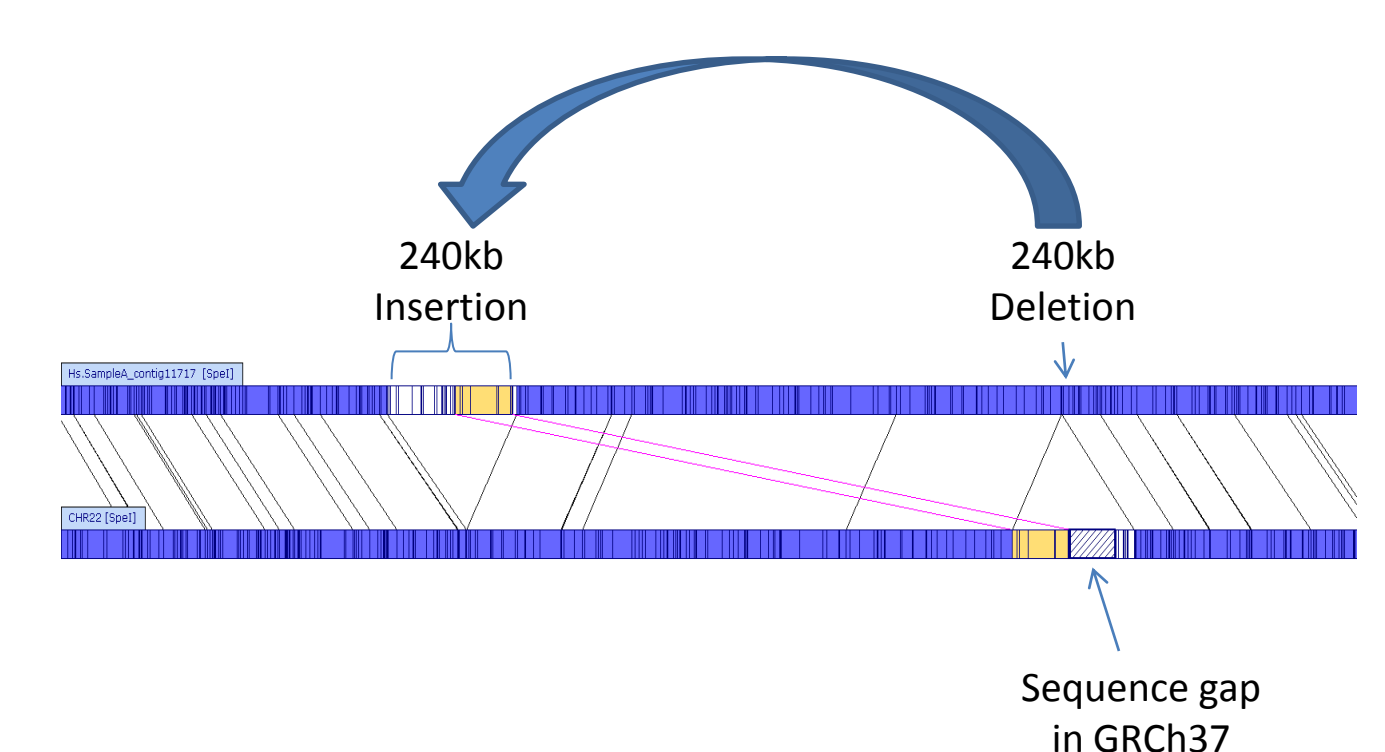


Using OpGen's comparative genomics analysis tool, the consensus maps are then used to detect structural variations (SVs) between different genomes. The following are a few examples of SVs between this genome and GRCh37 reference genome.

Multiple large inversions



A translocation in chromosome 22



Summary

1. The Argus Whole Genome Mapping System is capable of rapidly generating a large amount of high quality single molecule mapping data to support analysis of human genomes.
2. The Genome Builder software package provides a powerful tool that effectively and accurately bridges sequence scaffolds separated by gaps up to 200 kb in large complex genomes.
3. Whole Genome Mapping has shown great promise in depicting genomic architecture and detecting structural variations in human genomes.