Human Chromosome Explorer™ for HD-Mapping™

Structural variation analysis using electronic whole-genome mapping data

Overview

Nabsys collaborated with Hitachi High-Tech to develop Human Chromosome Explorer, a cloud-based analytical pipeline for high quality *de novo* assemblies of whole-genome maps and accurate structural variation calls visualized through a web browser.

INTRODUCTION

Human Chromosome Explorer (HCE) is an advanced, cloud-based, whole-genome assembly and structural variation (SV) analysis pipeline for the Nabsys HD-Mapping platform (Figure 1). Developed by Hitachi High-Tech in collaboration with Nabsys, HCE uses Google Cloud's High-Performance Computing (HPC) infrastructure to create haplotype-aware human whole-genome map assemblies and perform SV analysis with results visualized and reported through a web browser.

Figure 1: Nabsys HD-Mapping platform for human whole-genome SV analysis.

HD-MAPPING & SV ANALYSIS

The Nabsys HD-Mapping platform generates long single molecule reads for human whole-genome analysis of variants from approximately 300bp in size up to large chromosomal rearrangements. To do this high molecular weight (HMW) DNA is isolated, and the molecules are labelled at known recognition sites. Single DNA molecules are electrophoretically translocated through a solid-state nanochannel where the labels are electronically detected by changes in resistance caused by the analyte.

The results are single molecule maps of each DNA molecule with the location of, and distance between, each label. The single molecule maps are assembled by HCE into map contigs containing all labeled locations within the sample and then aligned against a reference genome map. Shifts in the loci of the labels are indicative of shifts in the genome that serve as the basis for SV analysis. For more details on HD-Mapping see [*"Nabsys HD-Mapping: A novel system for human whole-genome structural variation analysis."*]

FROM MOLECULE DETECTION TO SV CALL

Using HD-Mapping's onboard field-programmable gate array (FPGA), data is signal processed in real time and the location of the labels determined. The signal-processed reads are transferred to the cloud for SV analysis.

Upon uploading the signal-processed files to the cloud, the HCE analysis pipeline (Figure 2) extracts metadata statistics from the run to automatically tune assembly parameters. The maps are extracted and assembled into map contigs. High quality long-reads are identified and serve as the seeds from which to grow the map contigs through a process of iterative comparisons of smaller reads throughout the dataset. Unusable single-molecule maps are filtered out, eliminating the impact of these errant maps on assembly. This reduces the downstream computational burden

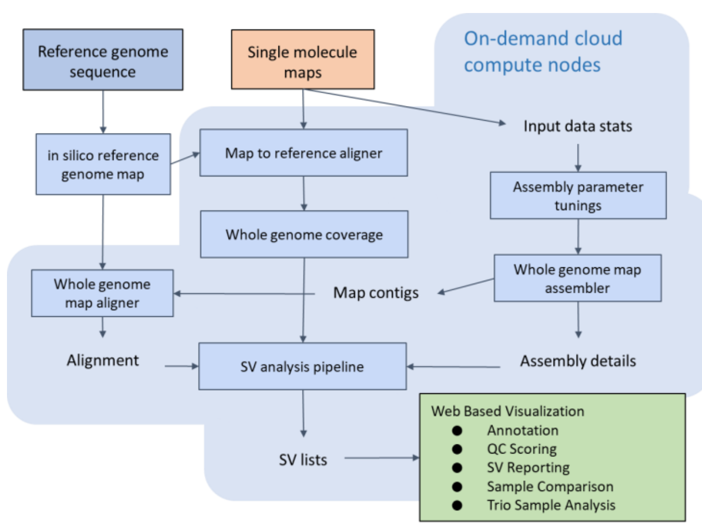

Figure 2: The HCE analysis pipeline uses Google Cloud's HPC environment to assemble HD-Mapping reads and perform human whole-genome SV analysis.

and allows for more of the quality reads to be used to create longer, and ultimately fewer contigs, in the final assembly. The assembly process is accomplished without seeding from the reference which allows *de novo* detection of SVs and decreases the incidence of false negatives due to reference bias.

Before SV analysis can begin, a comparable reference needs to be created. To accomplish this, a sequence-based reference genome is labelled, *in silico,* using the same sequence specific recognition sites as the sample to generate a reference genome map. The map contigs are aligned to the reference genome map. This alignment along with the assembly statistics are loaded into the SV analysis pipeline. SVs are then identified by comparing the observed label pattern of the map contigs with the expected pattern of the reference genome map. If label sites are added or removed, the size of an interval between two label sites has measurably changed, or there are changes to the loci of an interval compared to the reference it represents structural changes in the genome. For example, a region with a loss of genetic material between two labels on the sample creates a size difference in the interval compared to the expected interval on the reference and is called as a deletion. The size difference between the two intervals correlates to the size of the deletion. Conversely, this same principle is applied to insertions (Figure 3).

Once SVs have been identified, the list is checked against regions of the genome that contain segmental duplications and known gaps. Any SVs found in either case are reported as belonging to these regions. The list is then checked for zygosity, a QC score is assigned, and the SVs are genomically annotated. In parallel to the above assembly-based pipeline, unfiltered single molecule maps are aligned directly to the *in silico* labeled reference genome map. The number of maps aligned to a specific location within a chromosome can be quantified to determine copy number. The resulting analysis of both pipelines is visualized through HCE in a web browser where further filtering and analysis can be done.

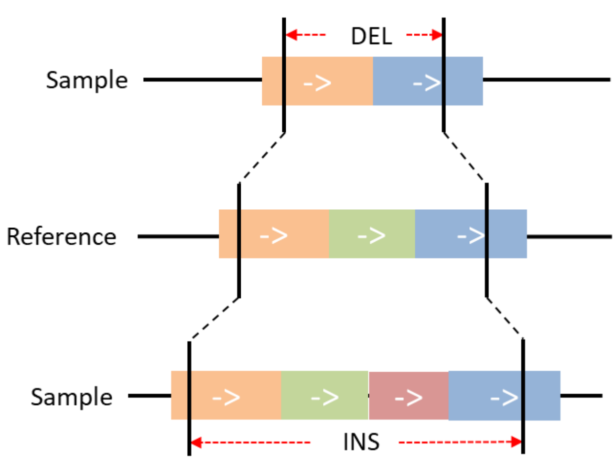

Figure 3: Examples of a deletion (top) and insertion (bottom) between label locations.

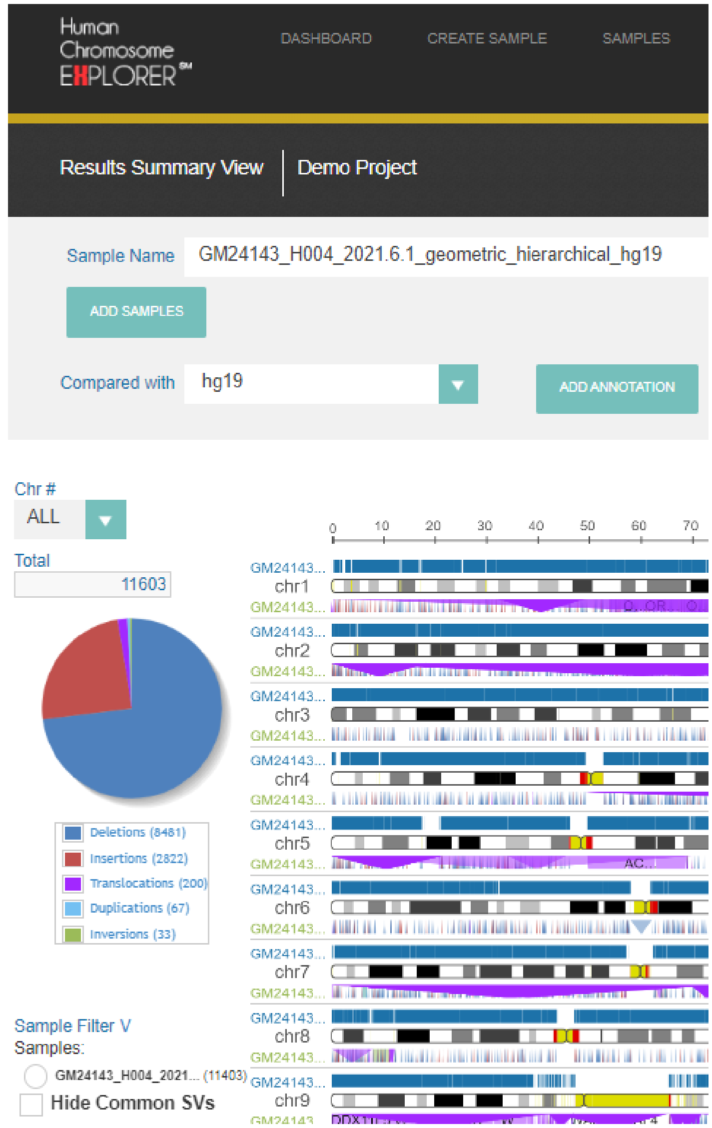

Figure 4: Screenshot of the Result Summary View in HCE.

HUMAN CHROMOSOME EXPLORER

HCE's SV analysis pipeline takes advantage of the power of the cloud. Google Cloud's HPC environment allows HCE to dynamically spin up the optimal number of compute nodes to speed up iterative tasks and enable complex parallel processing of human whole-genome mapping datasets. The results are high quality *de novo* assemblies and accurate SV calls without the use of compression which could result in data loss. Visualization on HCE is web-based through a simple user interface supported on Google Chrome or Microsoft Edge.

Once a user is logged in, a sample and reference genome are selected, and the analysis is initiated. In addition to automatically tuned parameters, optional parameters can be set by the user prior to triggering a run.

HCE has several visualization and reporting tools to navigate through SV analysis results. The Results Summary View (Figure 4) is the analysis dashboard. From here users can launch a Circos Plot viewer (Figure 5) or navigate the current view which displays SV type and count in a pie chart and a global view of all chromosomes. Map coverage, gene annotations, and SV locations are displayed. Users can zoom into a location of interest on any chromosome or filter the results by location, SV size, SV type, zygosity, or gene name, etc. Hovering at any location with display detailed information about that location including base pair location, any known genes, SV type, SV size, confidence, etc. Additional samples can be added to the view and external BED files can be imported to further annotate the samples.

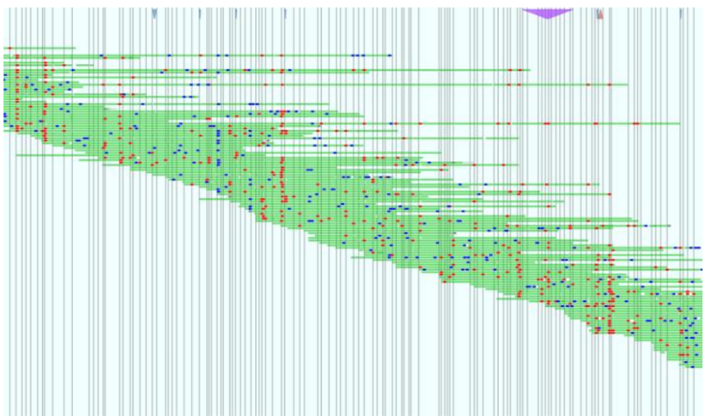

Figure 7: HCE's Assembly Viewer showing the depth, coverage and label locations of map contig assemblies.

Clicking on any location on a chromosome will launch the Alignment Viewer (Figure 6). This viewer shows all map contigs aligned to the reference of the selected chromosome. This interactive viewer allows users to drag and zoom to any location or jump to a particular location by entering the corresponding base pair designation. Hovering over any location on map contig will display size, location, molecule depth, selected interval size, etc.

Clicking on any location in the Alignment Viewer will launch the Assembly Viewer (Figure 7) where label locations, molecule

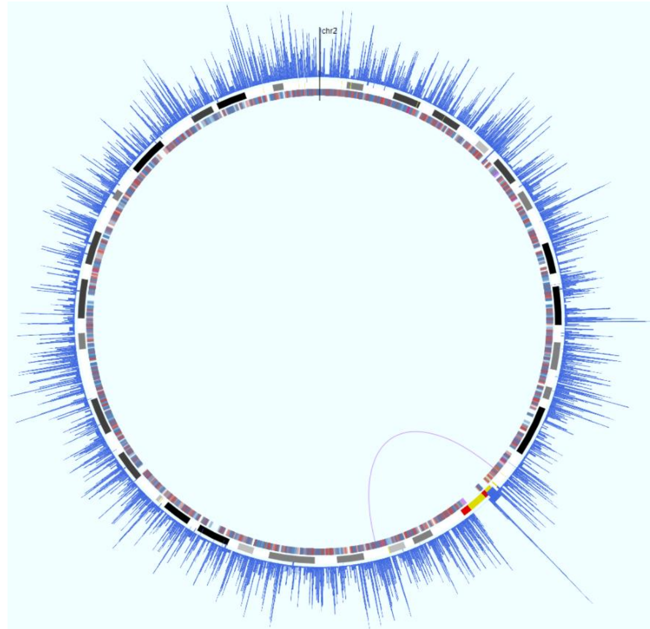

Figure 5: HCE's Circos Plot viewer showing a translocation on chromosome 2.

orientation, and alignments are displayed. As in the other views, hovering over a location will display detailed information about that location including molecule depth and coverage, molecule length, map contig ID, etc.

In the Results Summary View users can access additional reporting functions. The Stats view provides detailed statistics about the input data, assembly, and detected SVs. Input data stats include total length of genome coverage, median, mean, minimum and maximum size of molecules and intervals (distance between two labels.) Assembly stats include number, median, mean, and minimum and maximum length of map contigs. Additionally, a number of assembly stats are reported for each chromosome. SV stats include type, size, frequency, and chromosome location. A separate SV List report contains SV ID, chromosome, start position, end position, length, type, zygosity, confidence score, etc. Finally, a robust export wizard supports multiple file formats and report types.

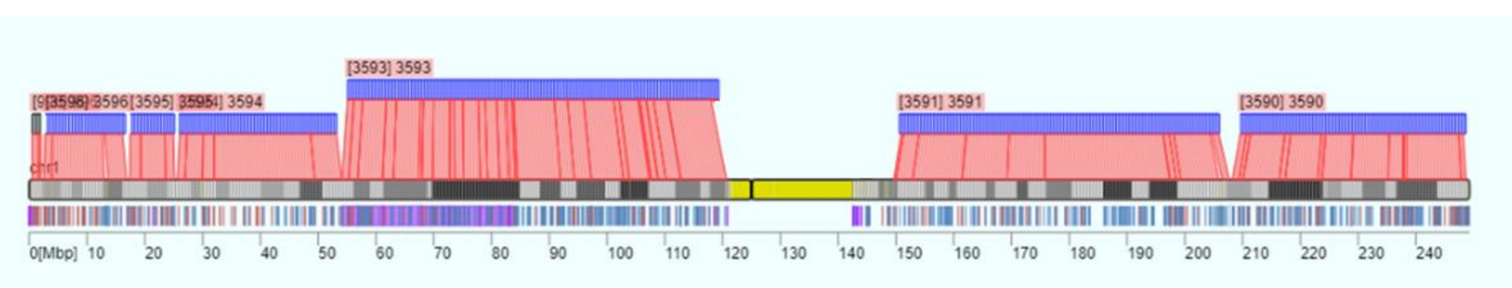

Figure 6: HCE's Alignment Viewer visualizes map contigs alignments to the reference genome map.

THE HCE SECURITY FRAMEWORK

Through its collaboration with Google Cloud (https://cloud.google.com), Hitachi High-Tech has developed a secure framework for the analysis and storage of genomic data.

Along with data encryption, the Google Cloud Platform provides a cloud infrastructure with robust built-in compliance and security features applied to the physical datacenter, and the network. Hitachi has further built upon this infrastructure to create a secure platform for genomic data called the HCE Security Framework. This framework includes the following:

- Protection against basic attacks (e.g., port scan, DDos, etc.) by the Google Cloud
- An additional security layer applied to the default Google Firewall, called the Web Application Firewall (WAF), to prevent more sophisticated attacks such as SQL injection attacks
- Storage (Database, Files) that is 100% Google Cloud based
- Google Cloud's full data encryption

CONCLUSION

The Human Chromosome Explorer analysis pipeline powered by the Google Cloud platform provides the computational power and secure framework to perform human whole-genome structural variation analysis on the high-resolution single-molecule reads generated by the Nabsys HD-Mapping system. Automatically tuned assembly and SV calling parameters and a simple UI accessed through a web browser make the pipeline accessible to all end users. SV analysis results can be visualized or exported for tertiary analysis. HCE when combined with HD-Mapping offers a cost-effective solution for human whole-genome structural variation analysis.

LEARN MORE

To learn more about Nabsys HD-Mapping visit https://nabsys.com