What is the BioGraph™ Format?

The BioGraph Format is a method for storing NGS read data using an extension of the Burrows Wheeler Transform to allow for multiple paths. Effectively, this creates a read overlap assembly of all the reads. All the read data are retained in this format. Further, a specialized index allows for very rapid search of the data, not just those reads that align to the reference.

The BioGraph Format is created in two steps using the reads in a BAM file:

- Error correction
- Construction of the BioGraph Format.

Error Correction

Error correction corrects the reads using a deBruijn graph of the 30-mers in the read set. Those 30-mers that occur three times or less are corrected to match the k-mer deBruijn graph with the fewest changes in base pairs (A*search). By reducing the number of paths due to error, the total size of each file remains small. The number times that a k-mer appears before it is corrected can be adjusted.

Some reads are not correctable because they are due to biological contamination. These reads are discarded.

Construction of the BioGraph Format

Once the reads are error corrected, the BioGraph Format is constructed. In practice, two files are created. The first is effectively the read overlap graph of all the reads in the read set. The second file is the path through that graph for the individual. As the BioGraph Format for individuals are merged together, files are created for each individual to indicate their path through the combined graph of all the reads.

Once the read overlap graph is constructed, a natural coordinate system is produced. In this, every location of every individual has a unique location. Regions that are common across individuals have identical coordinates. An individual is fully defined by the full set of coordinates that are present. Comparisons between individuals and groups can then be performed rapidly by comparing across this coordinate system.

References are added as a path through the graph, and it is possible for any number of references to be added to the graph. In this way, for any sequence of interest, it is possible to identify where it is located on any reference, for example GRCh37 as well as GRCh38. Furthermore, if there is a variation from the reference, it is possible to observe where that variation occurs across the coordinate system for different references.

File Size, Import/Merge Time, and Query Metrics

The following measurements were made using 30x BAM files generated using an Illumina HiSeq X. Computation was performed on a 32-core machine with 64GB of RAM.

File Type	File Size
BioGraph for one individual	23GB
BioGraph for 50 genomes	342GB
Effective file size per genome at scale	6.8GB

Computation Type	Time
Convert from BAM to BioGraph	10 hours
Merge 2 BioGraph genomes together	1 hour
Merge 50 BioGraph genomes together	25 hours (1/2 hour per genome)

Number of Genomes	1 – 100 Variants	100,000 Variants	1,000,000 Variants
1 genome	Instant	1 second	10 seconds
100 genomes	5 minutes	7 minutes	22 minutes
10,000 genomes	8 hours	8 hours	37 hours
100,000 genomes	3 days	5 days	15 days