NA12878 BRCA1 and BRCA2 Variants
Crystal Genetics, Inc.

www.crystalgenetics.com

Summary

This is a comparison of the variants (SNVs, insertions, deletions and block substitutions) of NA12878 for BRCA1 and BRCA2 genes between the calls made by Crystal Genetics versus Illumina's Platinum Genomes V7.1.0 (herein referred to as the Benchmark). All the data used for the Benchmark can be found at Illumina's PlatinumGenome's website: http://www.illumina.com/platinumgenomes .

- In our opinion, the Benchmark provides one of the highest quality (in terms of accuracy and comprehensiveness combined) public whole-genome variation calls for NA12878 in the world. In order to achieve its accuracy, this benchmark has utilized several leading variant callers and pedigree information from the three-generation family tree of NA12878 (see below for details).

- We also believe that (in view of the interrogated genes presented here) the variant calls made by Crystal Genetics are highly competitive with those made by the Benchmark.

- The intention of this comparison is not to have a head-to-head comparison between Crystal Genetics and the Benchmark as methods, as they are designed for different purposes:
    - Crystal Genetics focuses on specific genes and targets clinical-grade accuracy for those genes.
    - The Benchmark focuses on providing the baseline whole-genome variants of an individual, and does so by pooling variations from multiple constituent algorithms across the pedigree corresponding to that individual. In other words, the objective of the Benchmark is to boost the accuracy and comprehensiveness by utilizing related genomes and different algorithms.

Results

The results are summarized in the associated Excel files:

- The variants that are not highlighted represent identical calls between Crystal Genetics and the Benchmark.
- The variants that are highlighted in green represent consistent calls, based on Crystal's visual analysis of the reads supporting the variants.
- The variants that are highlighted in red represent inconsistent calls, based on Crystal's visual analysis of the reads supporting the variants.
- The variants that are highlighted in yellow represent equivalent and consistent calls, based on Crystal's visual analysis of the reads supporting the variants.

- The variants that are highlighted in grey represent hard-to-assess calls, based on Crystal's visual analysis of the reads supporting the variants.

Please note that the consistent/inconsistent designation does not mean to reflect a correct/incorrect state for the calls. The consistent and inconsistent designation is meant to reflect the state of the calls as compared to Crystal's visual checking of the reads supporting the corresponding variants. Please also note that the visual analysis is not meant to be a replacement for a confirmation assay, which would be the ultimate way of validating the quality.

Raw Data

- Both Crystal's and the Benchmark's call sets are based on the same exact raw whole-genome sequencing (WGS) data, provided by Illumina in http://www.illumina.com/platinumgenomes. This data comprises a 50x human genome sequenced with Illumina's HiSeq2000 system.
    - Crystal Genetics does not use any other (implicit or explicit) raw data.
    - The Benchmark implicitly uses other raw data, by including variants from Complete Genomics, which are called using a different set of reads (provided by Complete Genomics).

Leveraging Family Structure

- Crystal Genetics uses only the data for NA12878 and does not rely on access to any family structure, as it is targeted for clinical genomes, where (often) only the patient's data is available.
- The Benchmark uses a 17-member family-tree (CEPH Pedigree 1463). This family-tree is composed of NA12878 and the associated three-generation genomes (of the parents, spouse, spouse's parents, and 11 children).

Variant Calls

Both reported variants calls have used Build 37 of Human Reference Genome for reporting.

- Crystal Genetics' variant calls are obtained using its own proprietary algorithms --from raw reads (FASTQ) to variants (VCF)-- and without utilizing any consensus data or bioinformatics/other databases (e.g., dbSNP). It uses only the read data from one sequencing platform (Illumina) from that individual.
- The Benchmark's variant calls are based on the consensus of different open-source and proprietary algorithms and sequencing platforms. The benchmark lists the following algorithms as its constituents:
    1. CGI (Complete Genomics, Inc.)
    2. BWA-FreeBayes (Boston College)
    3. BWA-Platypus (The Wellcome Trust Centre for Human Genetics)
    4. BWA_GATK3 (Broad Institute)
    5. Cortex (The Wellcome Trust Centre for Human Genetics)
    6. Isaac2 (Illumina, Inc.)

Please note that the results reported in this document are not intended as a direct comparison between Crystal Genetics' algorithms and any of the above algorithms. The reported results are only intended for the combination effect of these 6 algorithms as used by the Benchmark. In other words, it is possible that if one uses a different version of these algorithms (than the one used in the Benchmark) or a different set of parameters in running these algorithms (than the set of the parameters used in the Benchmark), or if the algorithms' outputs are not filtered based on the pedigree constraint, the results may be widely different.

Reports

In order to provide a comprehensive view for the variant calls, a union of the reported positions in Crystal and the Benchmark was made. Subsequently, for each method, the reported loci were loaded from the VCF files into the corresponding positions.

For each of Crystal and the Benchmark, a column (named varcall) was added to show if the corresponding variant was found in the VCF file or not. If no variant was found in the VCF file, the other columns were filled with a dash. This was in attempt to line-up the variants of Crystal and the Benchmark. Please note that a dash should not be interpreted as no data, as it may, for instance, contain ref/ref calls or variants which are filtered out in one of the call sets.

Low-quality Calls

- Crystal's calls that are reported are all considered high-confidence calls.
- The Benchmark may have categories of low-confidence calls –If the filter designation is not equal to PASS. In our analysis, the low-confidence calls were marked as no-calls, prior to comparison.

Acknowledgement

We would like to thank Illumina, and in particular Dr. Michael Eberle, for the insight into different versions of Platinum Genomes, and for the invaluable comments in reviewing this document and its associated data.

Questions

- If you have any questions, please feel free to reach us via the Contact Tab at our website: www.crystalgenetics.com