Fabric Genomics' Opal™ Clinical Variant Interpretation Platform Enables Rapid Whole Genome Analysis Turnaround in Under an Hour

Anthony Fejes[1], Björn Stade[1], Stephan Ritter[1], Anna Lewis[1], Edward Kiruluta[1], Martin Reese[1]
1. afejes@fabricgenomics.com, Fabric Genomics™ Inc., Oakland, CA

ABSTRACT

Opal™ Clinical for is well suited for use in clinical applications, supporting delivery of rapid diagnostic outcomes on whole genome data.

NGS-based diagnostics are poised to revolutionize the way clinical labs identify and interpret inherited or *de novo* diseases. However, the long sequencing and processing times make it difficult to incorporate into existing workflows. For diagnostics purposes, a raw sample must be sequenced, aligned to the reference genome, called for variant bases, annotated and interpreted.

While the speed and cost of sequencing 30x whole genomes, suitable for diagnostic purposes, continues to improve, accurate genome annotation and interpretation remains one of the most challenging and time consuming components of the end-to-end process. To solve this issue, Fabric Genomics™ has designed a platform that performs variant annotation for a whole genome in ~8 minutes. It provides an interface for interpretation that can yield immediate results for many diseases and enables deep genome interrogation and interpretation.

The quality of the annotations is ensured through robust regression testing. As each data source is upgraded, or source code is modified, data annotation tests are performed to ensure the accuracy of the annotations is not compromised. This ensures that the Opal™ Clinical annotation engine remains both up to date and accurate in its predictions.

We discuss here some of the benchmarks of our complete genome annotation platform, as well as some of its key features and design principles.

This annotation engine is currently in use widely, including in the UK's 100,000 Genomes Project, as well as Rady's Children Hospital, where they are launching a rapid genome service to return results for infants and children in the Neonatal Intensive Care Unit (NICU) and Pediatric Intensive Care Unit (PICU) in 24 hours.

The End-to-End Process for Clinical Reporting using Next Generation Sequencing data

For an individual with an undiagnosed genetic condition, whole genome sequencing is increasingly used in an attempt to identify the one or two variants (out of approximately 4 million) that are causative of the patient's phenotype. Rapid whole genome testing is now being offered in some hospitals in urgent cases such as infants in the NICU and PICU, where time is of the essence. There is a many-stepped path from taking a sample from the patient (and ideally both their parents) to a useful clinical report:

- **Sample Prep & Sequencing:** From the time the sample is collected, it needs to be prepared in the laboratory, before being queued for sequencing. The raw sequence reads are output in a FASTQ file.
- **Secondary Analysis:** alignment of raw sequence reads to the human reference genome (stored in a BAM file); bases that differ from the reference genome are identified, and the resulting variants are stored in a VCF (Variant Call Format) file.
- **Genome Annotation and Interpretation:**
  - *Automated stage:* variants in a VCF are further processed, including clean-up and annotation
  - *Human interpretation:* Variant Scientists, or other qualified individuals, interact with the results to produce a document that can be shared with the physician

Within Fabric Enterprise™, both secondary analysis as well as annotation and interpretation are offered as part of one seamless solution. Here, we focus on the automated annotation stage.

Genome Annotation and Interpretation

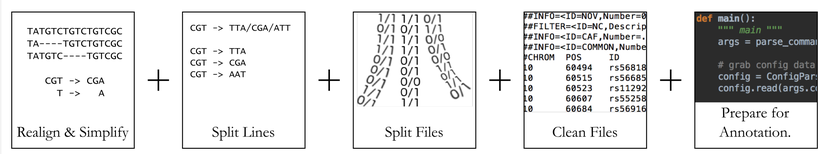
Traditionally, the most challenging component of the clinical analysis of Next Generation Sequencing data has been accurate annotation and interpretation, which is enabled by Opal™ Clinical.

- **Pre-processing:** The submitted file is validated, standardized and prepared for the annotation process
- **Annotation:** Each variant is pre-processed and annotated with over 100 individual data points
- **Clinical Report Generation:** Variants are assigned to inheritance modes, genes are algorithmically ranked incorporating phenotype information and are contextualized with information from a laboratory's previous cases

Subsequently, a case is ready for human interpretation within Opal™ Clinical's intuitive user interface, which is designed to enable Variant Scientists to use their time optimally. For example, using a Standard Operation Protocol that restricted human analysis time to 3 hours, a return of results in 45% of cases was achieved.

See Poster #411—Increased Yield of Clinically Relevant Candidates in the UK 100,000 Genomes Project Using Opal™ Clinical for hereditary disease.

Pre-processing



Each variant submitted to the Opal™ Clinical annotation engine is processed through a series of steps that normalize all valid representations, ensuring that variants will be consistently and correctly matched with all of our databases and scoring algorithms. This is primarily accomplished by using the method described in Tan et al. (2015), which performs a consistent left-alignment. A second step is to subject each variant to a multi-step validation, to ensure that the data provided meets the necessary minimal format requirements for interpretation, including matching reference base calls to the official genome reference, ensuring valid genotypes - requirements of VCF formats 4.0 to 4.2.

VCF files containing rows with multiple alleles are also split into individual records, and files containing more than one sample are separated, ensuring that each sample can be individually interpreted and processed, or re-combined as necessary for family studies. Files containing all of the called variants (non-reference) for download, while no-calls and ref-calls are retained separately for use in family studies.

A similar pre-processing methodology is applied to structural variants, then the structural variants are passed through a subsequent dedicated analysis engine.

The full set of variants are then released for annotation.

Genome Annotation

The annotation engine is designed to be:

- **Comprehensive.** Each variant is marked up with over 100 data points, covering a wide range of useful metrics to assist with variant classification and prioritization. This includes properties such as variant consequence and functional annotation (VEP), allele frequency (ExAC, 1000 Genomes, EVS6500), bioinformatics scores (Omicia Score, SIFT, MutationTaster, VVP, CADD, etc.), splicing predictions (NNSplice, MaxEntSplice, GeneSplicer) and Literature evidence (OMIM, ClinVar, COSMIC, etc) as well as features such as alternate transcripts, and tentative variant prioritization.
- **Up-to-Date.** A new version of the annotation engine is released each month, enabling data sources to be kept up to date. Many data sources relevant to the Human Genome are expanding rapidly, necessitating frequent updates to ensure that the most recent knowledge is available during interpretation.
- **Accurate and Consistent.** Regardless of the upstream platform and variant caller applied to the sample, the annotation process guarantees that high quality, consistent annotations are generated every time. Each version of the annotation engine is tested to ensure that there are no surprises or undocumented changes along with each update.
- **Fast.** The Opal™ Clinical annotation engine was designed to ensure that annotations can be performed "on-demand". Compared to published benchmarks (Yen et al. Genome Medicine (2017) 9:7), we are able to completely annotate full genomes in under 10 minutes, between 150 to 900 times faster than other available programs. For urgent processing, a high-priority or "STAT" option is available, guaranteeing turn around times that make the platform useful in situations where high-speed clinical interpretation is required, like in Neonatal Intensive Care Units.

Sample	Time to annotate
Genome (4.2M variants)	~8 minutes
Exome (100,000 variants)	42 seconds
Panel (~1000 genes, 2000 variants)	5 seconds

Clinical Report Generation

Annotated variants are then contextualized to enable Clinical Interpretation:

- Variants are assigned inheritance modes that are consistent with their presence. (e.g. a variant that is present in the proband, but not the mother or father, is marked *de novo*).
- Genes are algorithmically ranked using VAAST[1], which integrates sequence conservation, genetic consequence, and allele frequency in a probabilistic framework.
- The proband's phenotype information (entered as Human Phenotype Ontology terms) is used to re-rank the VAAST results using an algorithm called Phevor.[2]
- Variants that have been previously classified within a laboratory are marked with their prior classification and classification date.

For whole genome data, the clinical report generation process takes 5-7 minutes in Opal™ Clinical. Then the data is accessible for interpretation within an intuitive user interface.

SUMMARY

Genome annotation is one of the key challenges for the clinical interpretation of Next Generation Sequencing data, where accuracy, speed and update frequency are all required. This is highly important in the NICU where speed is critical in guiding patient care for newborns and every minute counts from the child's admission to the clinical diagnosis. The Opal™ Clinical genome annotation solution can be used to process whole genomes in ~8 minutes in a framework that harmonizes variant representation and ensures data sources used are up-to-date, accurate and reliable.

The Opal™ Clinical annotation engine includes many publicly available data sources and bioinformatics applications. We would like to thank the many researchers and volunteers who have contributed to these projects.

Thanks to Mark Yandell and his group who have collaborated on projects that power the variant interpretation and analysis engines including VAAST and Phevor.

1. Using VAAST to identify disease-associated variants in next-generation sequencing data. Brett Kennedy, Zev Kronenberg, Hao Hu, Barry Moore, Steven Flygare, Martin G. Reese, Lynn B. Jorde, Mark Yandell, Chad Huff. Curr Protoc Hum Genet. 2014 Apr 24;81:6.14.1-6.14.25.

2. Phevor combines multiple biomedical ontologies for accurate identification of disease-causing alleles in single individuals and small nuclear families. Marc V. Singleton, Stephen L. Guthery, Karl V. Voelkerding, Karin Chen, Brett Kennedy, Rebecca L. Margraf, Jacob Durtschi, Karen Eilbeck, Martin G. Reese, Lynn B. Jorde, Chad D. Huff, Mark Yandell Am J Hum Genet. 2014 Apr 3;94(4):599-610. doi: 10.1016/j.ajhg.2014.03.010

FABRIC GENOMICS™

FORMERLY Omicia

Learn more at

www.fabricgenomics.com ■ info@fabricgenomics.com ■ 510.595.0800