Detecting somatic variants in acute myeloid leukemia tumor-only samples using the HIVE DRAGEN Pipeline

Sean D **Smith**[a], Michael **Colgan**[b], Luis V **Santana-Quintero**[a], Konstantinos **Karagiannis**[a]
a – Center for Biologics Evaluation and Research (CBER), Food and Drug Administration (FDA), Silver Spring, MD, USA
b – Center for Drug Evaluation and Research (CDER), Food and Drug Administration (FDA), Silver Spring, MD, USA

Abstract

Background: Identifying somatic mutations is an essential step in cancer studies and personalized cancer therapy. Often a matched normal (germline) sample is not available, presenting computational challenges to identifying somatic variants in tumor-only next-generation sequencing (NGS) samples. We developed a tumor-only analysis pipeline on the High-performance Integrated Virtual Environment (HIVE) and applied the pipeline to a tumor-only dataset of acute myeloid leukemia (AML) whole exome samples.

Methods: The tumor-only DNA pipeline was developed and optimized using whole genome sequencing data from a triple-negative breast cancer cell line (HCC1395), which included a robust somatic truth set. The optimized pipeline, consisting of FastP, DRAGEN, and the HIVE germline filter, was implemented in HIVE and applied to 15 AML whole exome sequencing samples. The germline filter included five population databases: dbSNP, gnomAD, ClinVar, 1000 Genomes Panel of Normals, and ESP.

Results: Our HIVE tumor-only DNA pipeline produced somatic mutation results in strong agreement with AML variants detected in previous tumor-normal studies. Focusing on AML-related genes, our results suggested that false positive somatic variants were uncommon relative to true variant predictions.

Conclusion: Using the sequencing pre-processing, variant calling, and filtration strategies implemented in the HIVE tumor-only DNA pipeline, somatic variant detection and analysis is feasible without the need for a matched normal sample.

Introduction

Identifying somatic mutations is an essential step in cancer studies and clinical applications. Often a matched normal (germline) sample is not available due to cost-constraints or unavailability in retrospective studies. For tumor-only next-generation sequencing samples, variant detection tools face a nearly impossible task of identifying a relatively small number of somatic variants, especially low mutation burden cancers such as AML, in a sea of germline variants. Techniques relying on information within the tumor sample, such as expected differences in allele fraction distributions of germline and somatic variants, have limited resolution power. Databases of germline and somatic variants have grown in the number of samples and breadth of samples across many populations. Many variant detection tools accept a germline variant database as input. However, a multitude of population/germline databases are available, presenting a challenging task for the researcher to optimize variant database utilization. We developed a tumor-only variant detection pipeline based on the Illumina DRAGEN variant calling platform and five variant databases and have implemented our pipeline in the user-friendly High-performance Integrated Environment (HIVE).

Materials and Methods

Publicly available whole exome sequencing raw data (FASTQ files) from a triple-negative breast cancer cell line (HCC1395) and 15 AML samples were downloaded from NCBI's SRA database (SRP162370) and EBI's ArrayExpress (E-MTAB-6299), respectively. After adapter and read quality-trimming/filtering (FastP), sequencing data was aligned (human – GRCh38) and processed through the HIVE-adapted Illumina DRAGEN (version 3.7.5) small variant pipeline. The DRAGEN call set was produced running DRAGEN with duplicate read marking, orientation bias, and systematic noise options and filtering to retain variants with SOMATIC and PASS designation. The systematic noise bed file (WES_TruSeq_IDT_hg38_v1.0_systematic_noise.bed) was downloaded from Illumina[1]. For the HIVE DRAGEN pipeline (Figure 1), DRAGEN variants with PASS designation were filtered to remove germline variants and artifacts using five variant databases: ClinVar[2], NHLBI Exome Sequencing Project (ESP)[3], 1000 Genomes Panel of Normals (1000G PoN, 1000g_pon.hg38.vcf)[4], dbSNP (version 154)[5], and gnomAD (version 3)[6]. ClinVar was filtered to retain benign or likely benign variants, dbSNP154 was filtered to retain variants with INFO field COMMON designation, and gnomAD (76156 whole genome sequencing samples) was filtered to retain variants with ≥0.003% population frequency.
Results were benchmarked against a somatic truth set (1160 SNVs overlapping WES target region) and further analyzed with a germline truth set for HCC1395[7]. For the 15 AML samples, results were benchmarked against tumor-normal somatic call sets[8] for the WES target region and a reduced set of 96 commonly mutated AML genes derived from the Brigham and Women's Hospital (BWH) Rapid Heme Panel[9].

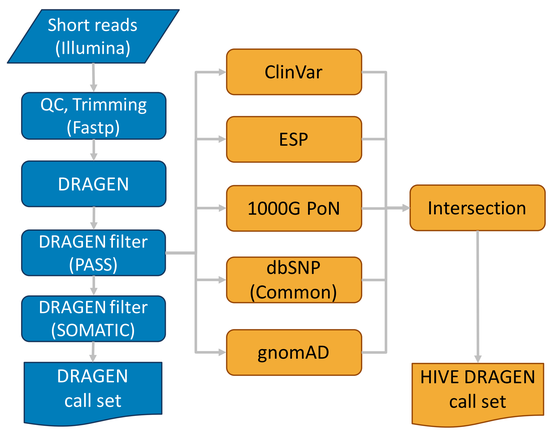

Figure 1. DRAGEN and HIVE DRAGEN pipelines.

Results and Discussion

DRAGEN's recommended somatic tumor-only pipeline (orientation bias, systematic noise, SOMATIC and PASS designation), called 7503 somatic SNVs, 543 true positives (46.8% sensitivity) and 6960 false positives (92.8% false discovery rate, FDR) when benchmarked using whole exome sequencing and a SNV somatic truth set for cancer cell line HCC1395 (Figure 2). For the HIVE DRAGEN tumor-only pipeline, we optimized thresholds and inclusion criteria for five variant databases (ClinVar, ESP, 1000G PoN, dbSNP, and gnomAD) to filter germline variants and artifacts from the DRAGEN output. To improve sensitivity, we included DRAGEN variants with non-SOMATIC designation (DRAGEN plus non-SOMATIC). This increased true positive somatic SNVs from 543 to 1077 (94.1% sensitivity) but ballooned false positives to 57788 (98.2% FDR). Germline filtering using the variant databases reduced false positives to 159 (13.3% FDR) while maintaining high sensitivity (89.4%, 1038 of 1160).

We attempted to recover true somatic variants that were removed by the HIVE DRAGEN pipeline filtration using the COSMIC Cancer Mutation Census (CMC). However, only three (1 true positive, 2 false positive) of the filtered variants were found in COSMIC CMC Tiers 1-3, and this option has not been included in the HIVE DRAGEN tumor-only pipeline.

We utilized the HCC1395 germline truth set to investigate false positives. For the DRAGEN, DRAGEN plus non-SOMATIC, and HIVE DRAGEN pipelines, 6115 of 6960 (87.9%), 53955 of 57788 (93.4%), and 139 of 159 (87.4%) false positives, respectively, were found in the germline truth set.

Higher throughput and lower sequencing costs have greatly increased the comprehensiveness of variant databases. One benefit may be improved variant detection capabilities using variant database filtration. Our results suggested significant gains (sensitivity improved from 46.8% to 89.4% and FDR reduced from 92.8% to 13.3%) by replacing DRAGEN SOMATIC filtering with the HIVE DRAGEN variant database filtering approach.

We further tested the HIVE DRAGEN pipeline using 15 AML tumor-only WES samples. AML has one of the lowest mutation rates (21 somatic SNVs/exome for our tumor-normal WES truth sets) among cancers[8], making AML one of the most challenging cancers for identifying somatic variants in tumor-only sequencing samples. Using published tumor-normal somatic variants as a truth set[8], the HIVE DRAGEN pipeline found 175 of 319 somatic SNVs (54.9% sensitivity) and had 1390 false positives (88.8% FDR) across the 15 AML samples (Figure 3). The average false positives per AML sample (92.7) was less than the false positives (159) for the breast cancer cell line. Restricting analysis to a set of 96 commonly mutated AML genes derived from the BWH Rapid Heme panel, FDR was reduced to 13.8% (4 false positives) while sensitivity was 45.5% (25 true positives).

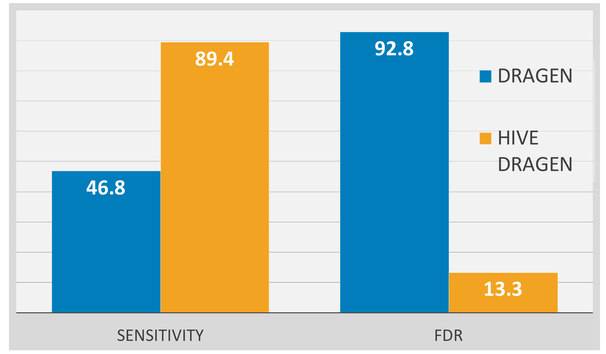

Figure 2. Comparison of somatic SNV call sets for tumor-only DRAGEN and HIVE DRAGEN pipelines. Call sets benchmarked using HCC1395 somatic truth set (1160 SNVs).

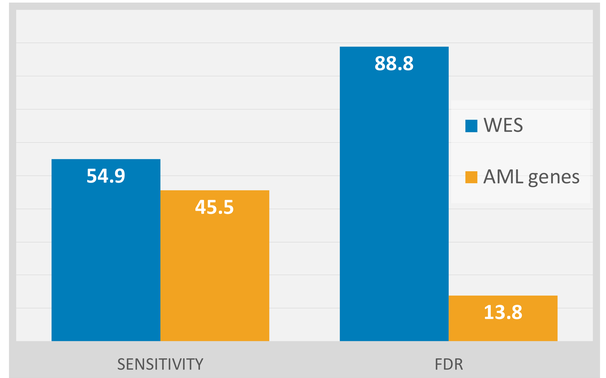

Figure 3. Comparison of somatic SNV AML call sets for the tumor-only HIVE DRAGEN pipelines. AML genes set corresponds to 96 genes targeted by the BWH Rapid Heme Panel.

Conclusion

Variant databases have become increasingly comprehensive enabling effective germline filtration strategies for tumor-only sequencing samples. We developed a tumor-only variant detection pipeline based on the Illumina DRAGEN variant calling platform and five variant databases, improving sensitivity from 47% to 89% and reducing FDR from 93% to 13% compared to DRAGEN alone on a breast cancer cell line. Low mutation burden cancers, such as AML, are particularly challenging for tumor-only sequencing analysis, however, by limiting our call set to commonly mutated AML genes, sensitivity was an average of 45% and FDR was 14% across the 15 AML samples. Our optimized tumor-only DRAGEN pipeline has been implemented in the user-friendly High-performance Integrated Environment (HIVE) and is easily accessible to the FDA community at https://scihive.fda.gov.

References

1. https://webdata.illumina.com/downloads/software/dragen/systematic-noise-baseline-collection-1.0.0.bin
2. https://ftp.ncbi.nlm.nih.gov/pub/clinvar/vcf_GRCh38/archive_2.0/2020/clinvar_20201003.vcf.gz
3. http://evs.gs.washington.edu/evs_bulk_data/ESP6500SI-V2-SSA137.GRCh38-liftover.snps_indels.vcf.tar.gz
4. https://console.cloud.google.com/storage/browser/gatk-best-practices/
5. https://ftp.ncbi.nih.gov/snp/
6. https://gnomad.broadinstitute.org/downloads
7. Fang LT, et al. (2020) *Biorxiv*, https://www.biorxiv.org/content/10.1101/625624v3
8. Quek L, et al. (2018) *Nature Medicine*, https://www.nature.com/articles/s41591-018-0115-6
9. Kluk MJ, et al. (2016) *J Mol Diagnostics*, https://www.sciencedirect.com/science/article/pii/S1525157816300241?via%3Dihub