Integration of Bina RAVE for genomic data analysis at the Center for Inherited Disease Research

K.N. Hetrick, B. Myers, S.M.L. Griffith, D. Snyder, K.F. Doheny, L. Watkins
Center for Inherited Disease Research (CIDR), IGM, JHU-SOM, Baltimore, MD

Introduction

The Center for Inherited Disease Research (CIDR) provides high quality genotyping and sequencing services and statistical genetics consultation to investigators working to discover genes that contribute to disease. With the ever increasing volume of NGS data being generated and the constantly evolving ways to analyze that data, software development, bioinformatics, and IT personnel can find themselves straining to maintain the levels of expertise required to support the high performance computing resources needed for providing consistently high levels of quality and efficiency along with fast turnaround times. There are a bevy of commercial platforms that aim to lessen this burden with a multitude of those aimed at uploading and analyzing data in the cloud. However due to privacy concerns with sample data, cloud-based solutions are not an option for all facilities. CIDR assessed the Bina RAVE (Read Alignment, Variant Calling and Expression) module (Bina Technologies, Redwood City, CA) deployed on a single Bina Rack local analysis appliance (4 node, 64 cores) with optimized workflows for whole genome/exome (WGS/WES) SNV/indel calling, WGS CNV calling, tumor/normal, and RNA analyses.

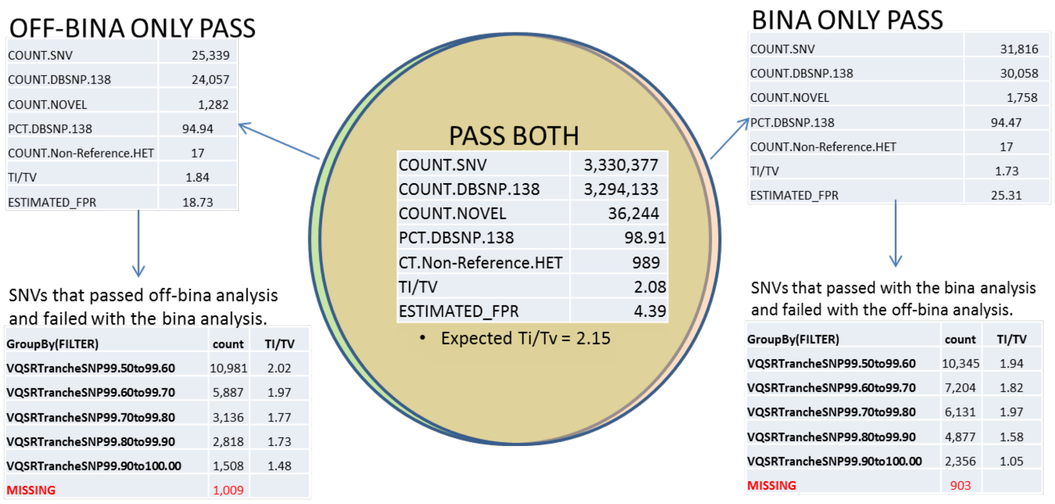
Comparison of WGS analyzed off and on appliance

We focused on comparing the results of a WGS sample sequenced on an Illumina® HiSeq™ 2000 at 38x depth analyzed on the Bina appliance using bwa mem 0.7.8 and GATK 3.1-1 (including HaplotypeCaller in GVCF mode) to an analysis generated outside of the appliance using the same workflow and software versions.

Good overlap between both analyses for SNV calls (Figure 1)

- 99.9% overlap for emitted SNV calls
- 98.3% overlap for SNVs passing VQSR
- 98.6% overlap for emitted indel calls (data not shown)

Figure 1 – Venn Diagram of SNVs that pass Variant Quality Score Recalibration (VQSR) filter for a WGS sample processed on the Bina appliance versus processed off the appliance.



Estimated_FPR (Estimated False Positive Rate):

((1-((ObservedTi/Tv – 0.5)/(Expected Ti/Tv -0.5)))*100

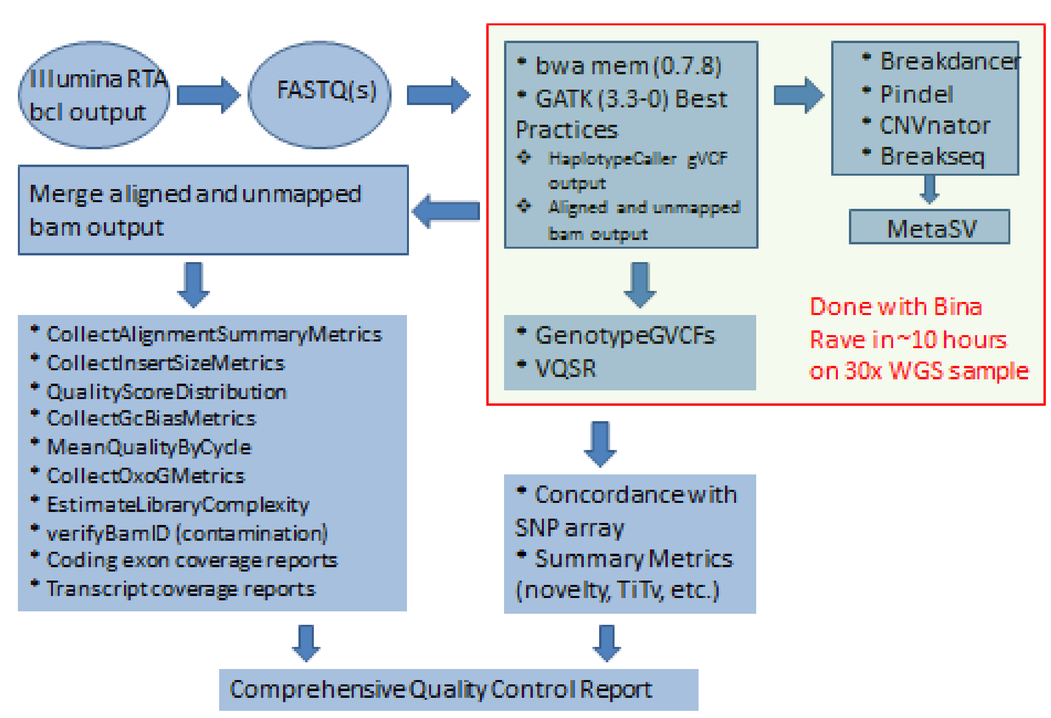
The wall clock time for the Bina RAVE analysis to perform the analysis was approximately 10 hours. Based on the above comparison and time to complete the analysis, we felt comfortable integrating the Bina RAVE analysis module with CIDRSeqSuite (see Poster 1630W; Myers, et al.) instead of developing a distributed WGS alignment and variant calling pipeline ourselves.

Integration of Bina RAVE module for WGS analysis

Integrated as part of CIDRSeqSuite with Bina API (Figure 2)

- bcl files are demultiplexed and converted into FASTQ
- json files are constructed to submit FASTQ files to Bina RAVE module (DNA workflow)
  - BAM files
  - GATK HaplotypeCaller gVCF
  - Structural Variant Analyses (Breakdancer, etc.)
- Bina DNA workflow monitored and after completion
  - Merge Bina bam output and submit QC metric reporting programs (e.g. CollectAlignmentSummaryMetrics, etc.)
  - Bina MDNA (GenotypeGVCFs and VQSR) json constructed and submitted to Bina appliance
    - Bina MDNA job monitored and other summary metrics programs (SNP array concordance, etc) submitted upon completion.

Figure 2 – Whole genome DNA sequencing analysis workflow integrating Bina RAVE module with CIDRSeqSuite



We are also interested in utilizing the Bina RAVE analysis module to perform the GATK joint calling workflow for our exome studies that involve hundreds to thousands of exomes. Using the gVCF output that comes from our own WES analysis pipeline, Bina was able to perform joint calling with GenotypeGVCFs and VQSR in two hours with close to 800 samples. Job submission just involves constructing a single json file listing all of the gVCF files. With this, we will not have to construct a distributed workflow to perform this within a comparable time frame saving programmer development time.

Conclusions and future work

- The Bina RAVA analysis module performs fast SNV/indel and SV analyses that compares favorably with analyses done outside the appliance
- Saves programmer development time. Allows easy job submission for lab personnel
- Investigate other Bina workflows (somatic, tumor/normal)