Overview of Variant Calling Workflow and Quality Metrics Obtained for New VCF Files on Whole Genome Sequences of ADNI Participants

Prepared and posted December 22, 2014
by

Jae Hoon Sul, PhD
Shamil Sunyaev, PhD
Robert C. Green, MD, MPH

Division of Genetics, Department of Medicine
Brigham and Women's Hospital, Broad Institute and Harvard Medical School

Introduction

Whole genome sequencing provided with support of the Brin-Wojcicki Foundation and the Alzheimer's Association was performed on 818 subjects from the ADNI Study by Illumina's non-CLIA laboratory at roughly 30-40x coverage in 2012 and 2013, see details posted here. The resulting variant call files (VCFs) generated by Illumina using CASAVA software were posted to the ADNI-LONI site in September, 2013 for distribution to approved investigators. Copies of the raw data in the form of binary alignment/map (BAM) files were stored at LONI, at Indiana University and at the Broad Institute and have been available from LONI for download by investigators who wish to send a hard drive as per the procedure described here and here.

In 2014 the Broad Institute donated in-kind effort, storage and compute time to take the BAM files back to their rawest form that includes quality data (FASTQ format), then re-aligned and re-called the data to produce new VCF files from the raw data using Broad GATK "best practices" as described below. The major advantage of the new VCF files is better accuracy of all calls, especially indel calls, due to the joint calling procedures, and because other sequences being performed in the ADSP project will be using GATK joint calling.

The variant calling workflow description below is composed of two main sections that were performed sequentially:

- Pre-processing and realignment: from raw DNAseq sequence reads (FASTQ files) to analysis-ready reads (BAM files)
- Variant calling: from reads (BAM files) to variants (VCF files)

Pre-Processing

The data generated by the sequencers were put through several pre-processing steps to make it suitable for variant calling analysis. This section describes the necessary pre-processing steps that were performed to prepare data for analysis, starting with FASTQ files, which were recovered from Illumina delivered BAM files, and ending in an analysis-ready BAM file after re-aligning.

These steps were performed in the following order:

1. Mapping and Marking Duplicates
2. Local Realignment Around Indels
3. Base Quality Score Recalibration (BQSR)

The sequence reads were first mapped to the reference using BWA-mem to produce a file in SAM/BAM format sorted by coordinates. The next step was to mark duplicates so that uninformative duplicate reads are not be counted as additional evidence for or against a putative variant. The duplicate marking process (sometimes called “de-dupping” in bioinformatics) identified these reads as such so that the GATK tools knew they should ignore them. The GATK version used was GenomeAnalysisTK-3.1-144-g00f68a3.jar.

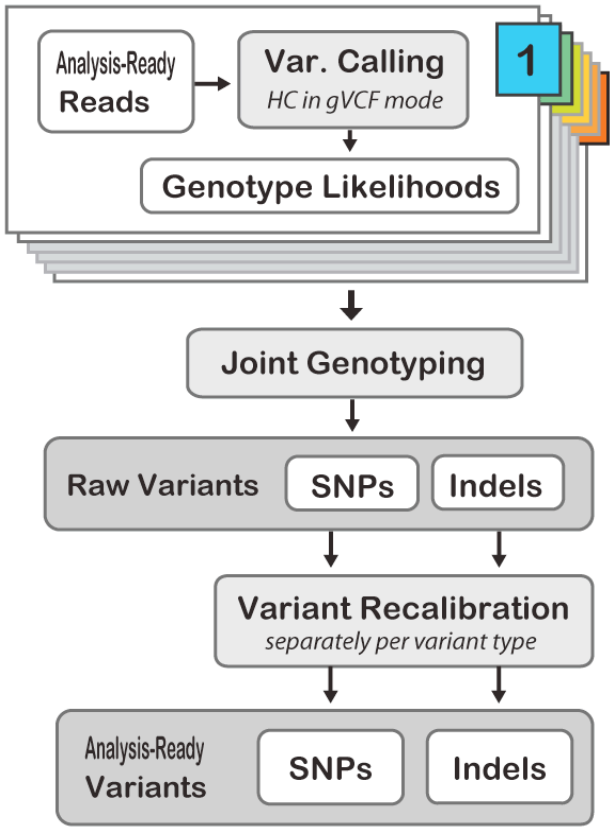
Next, local realignment was performed around indels, because the algorithms that are used in the initial mapping step tend to produce various types of artifacts. For example, reads that align on the edges of indels often get mapped with mismatching bases that might look like evidence for SNPs, but are actually mapping artifacts. The realignment process identifies the most consistent placement of the reads relative to the indel in order to clean up these artifacts. This realignment occurred in two steps: first the program identified intervals that needed to be realigned, then in the second step it determined the optimal consensus sequence and performed the actual realignment of reads.

Finally, base quality scores were recalibrated because variant calling algorithms rely heavily on the quality scores assigned to the individual base calls in each sequence read. These scores are per-base estimates of error emitted by the sequencing machines. Recalibration is necessary because the quality scores produced by the machines are subject to various sources of systematic error, leading to over- or under-estimated base quality scores in the data. Base quality score recalibration is a process in which machine learning is applied to model these errors empirically and adjust the quality scores accordingly. This keeps base qualities more accurate, which in turn improves the accuracy of the variant calls. The base recalibration process involved two key steps: first the program built a model of covariation based on the data and a set of known variants, then it adjusted the base quality scores in the data based on the model.

Variant Discovery

Once the data were pre-processed as described above, the next step was variant discovery, i.e. the identification of sites where the data displays variation relative to the reference genome, and calculation of genotypes for each sample at that site. Because some of the variation observed is always caused by mapping and sequencing artifacts, the greatest challenge here is to balance the need for

sensitivity (to minimize false negatives, i.e. failing to identify real variants) vs. specificity (to minimize false positives, i.e. failing to reject artifacts). It is very difficult to reconcile these objectives in a single step, so instead the variant discovery process is decomposed into separate steps: **variant calling** (performed per-sample), **joint genotyping** (performed per-cohort) and **variant filtering** (also performed per-cohort). The first two steps were designed to maximize sensitivity, while the filtering step aims to deliver a level of specificity that can be customized for each project.



1. Variant calling

The GATK HaplotypeCaller was run on each sample's BAM file(s) (if a sample's data was spread over more than one BAM, then they were passed all in together) to create single-sample gVCFs. Documentation reference for the GATK HaplotypeCaller is provided here.

2. Data aggregation step

Since we had more than a few hundred ADNI samples, we ran CombineGVCFs on batches of ~200 gVCFs to hierarchically merge them into a single gVCF. This made the next step more tractable and reflected that the processing bottleneck was with the number of input files and not the number of samples in those files. Documentation reference for the GATK CombineGVCFs is provided here.

3. Joint genotyping with all available samples

GenotypeGVCFs were run on all of the outputs together from step 2 to create a set of raw SNP and indel calls. Documentation reference for the GATK GenotypeGVCFs is provided here.

4. Variant recalibration

We used a machine learning method to assign a well-calibrated probability to each variant call in a raw call set. We then used this variant quality score in the second step to filter the raw call set, thus produced a subset of calls with our desired level of quality, fine-tuned to balance specificity and sensitivity. Documentation references for the GATK VariantRecalibrator and the ApplyRecalibration are provided here and here, respectively.

Additional Processing and Generation of Quality Metrics

The laboratory of Shamil Sunyaev, PhD at Brigham and Women's Hospital and Harvard Medical School reviewed the processing described above and provided additional quality metrics as described below.

After dropping one individual for quality control (QC) issues during BAM re-processing and dropping 9 individuals who were considered to have substandard consent, of the VCF files, 808 individuals were examined. Variants that failed Variant Quality Score Recalibration in the GATK pipeline were excluded, and genotypes whose genotype quality (GQ) scores ≤ 20 were set to as recommended by the GATK team. After removing monomorphic variants, a total of 44,535,780 variants are present in the revised VCF file. These are the VCF files that were uploaded to LONI.

Additional QC Procedures

We generated a collection of statistics on genetic variants discovered through sequencing and annotated them using MapSNPs (MapSNPs is part of the PolyPhen2 software suite). Table 1 shows these statistics, such as the number of SNVs and indels discovered, the number and percentage of novel SNVs and indels, and the number of singletons. Table 2 displays the results of annotation using MapSNPs and shows the number of missense, nonsense, and synonymous variants, transition/transversion (Ti/Tv) ratio, and CpG and non-CpG ratio. The tables are generated with reference to dbSNP Build 135.

Table 1: Various statistics on variants discovered by WGS	
# of SNVs	38,443,567
# of SNVs in dbSNP	20,833,975
% of SNVs in dbSNP	54.19%
# of novel SNVs	17,609,592
% of novel SNVs	45.81%
Ti/Tv ratio of SNVs in dbSNP	2.2222
Ti/Tv ratio of novel SNVs	2.0217
# of indels	6,092,213
# of indels in dbSNP	2,049,034
% of indels in dbSNP	33.63%
# of novel indels	4,043,179
% of novel indels	66.37%
# of multi-allelic SNVs	254,079
% of multi-allelic SNVs	0.66%
# of multi-allelic SNVs in dbSNP	193,212
% of multi-allelic SNVs in dbSNP	0.93%
# of singleton SNVs	17,983,148
% of singleton SNVs	46.78%
# of singleton indels	1,911,812
% of singleton indels	31.38%

Table 2: MapSNPs annotation of variants	
All SNVs	38,443,567
Annotated SNVs	14,988,341 (38.99%)
missense	212,063 (1.41%)
nonsense	4,263 (0.03%)
coding-synon	146,471 (0.98%)
intron	14,232,400 (94.96%)
utr-3	327,009 (2.18%)
utr-5	66,135 (0.44%)
missense/coding-synon	1.45
nonsense/coding-synon	0.03
CpG	2,920,001
Non-CpG	10,247,112
CpG/Non-CpG	0.28
Ti (all)	10,301,900
Tv (all)	4,686,441
Ti/Tv (all)	2.2
Ti/Tv (coding)	2.97
Ti/Tv in CpG (all)	8.61
Ti/Tv in Non-CpG (all)	1.67
Ti/Tv in CpG (coding)	7.6
Ti/Tv in Non-CpG (coding)	2.03

We also performed basic GWAS QC where we estimated identity by descent (IBD) estimates called $\hat{\pi}$ to detect related individuals, performed principal components analysis (PCA) of the 808 individuals, and checked genotype concordance between WGS and Omni 2.5M microarray data that had been previously collected on the same individuals. We identified five pairs of individuals with $\hat{\pi}$ between 0.4 and 0.6, which indicates that each of these pairs is first-degree relatives. These five pairs of related individuals are as follows: 024_S_2239 and 024_S_4084, 067_S_0056 and 067_S_0059, 021_S_0159 and 137_S_4466, 023_S_0058 and 023_S_4035, and 031_S_4032 and 031_S_4203.

Figure 1 shows a PCA plot using 1000 genomes as the reference panel. The majority of individuals are Europeans while there are several African Americans and Asians. Figure 2 shows a histogram of genotype concordance rate among the 808 individuals. All individuals have at least 99.88% genotype concordance rate between WGS and Omni 2.5M microarray data previously performed on the same individuals.

Figure 1. PCA plot of ADNI and 1KG using the top 2 PCs

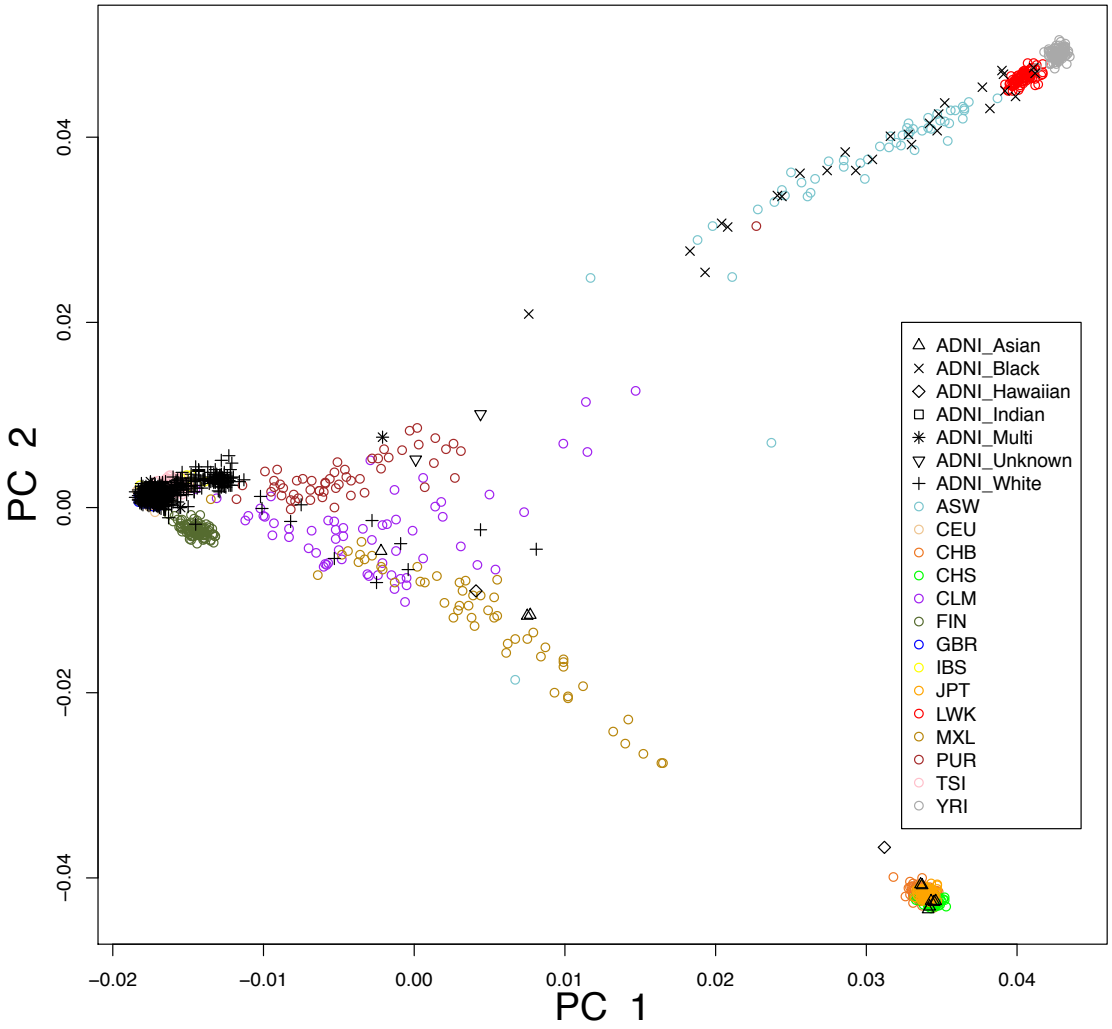


Figure 2. Genotype concordance of ADNI individuals between WGS and Omni 2.5M microarray data