Evaluation of variant calling from thousands of low pass WGS data using GATK haplotype caller

Hua Ling, Kurt Hetrick, Peng Zhang, Elizabeth Pugh, Jane Romm, Kimberly Doheny
Center for Inherited Disease Research (CIDR), Johns Hopkins University, Baltimore, MD 21224

Introduction

Low pass WGS (~2-8x) on large numbers of samples has become an attractive strategy in genetic studies of complex traits. Given the same amount of yield, it provides more power to detect disease associated variants than deep sequencing (30x) a small number of samples. It can also be used to build reference panel for imputing additional samples to further boost power. GATK haplotype caller (HC) now allows us to do joint calling and analysis on multiple WES samples which was computational prohibitive. HC is superior to Unified Genotyper (UG) in that it not only provides more accurate variant calls (especially for INDELs), but offers more flexibility by allowing for adding more samples in later stage without re-processing the cohort.

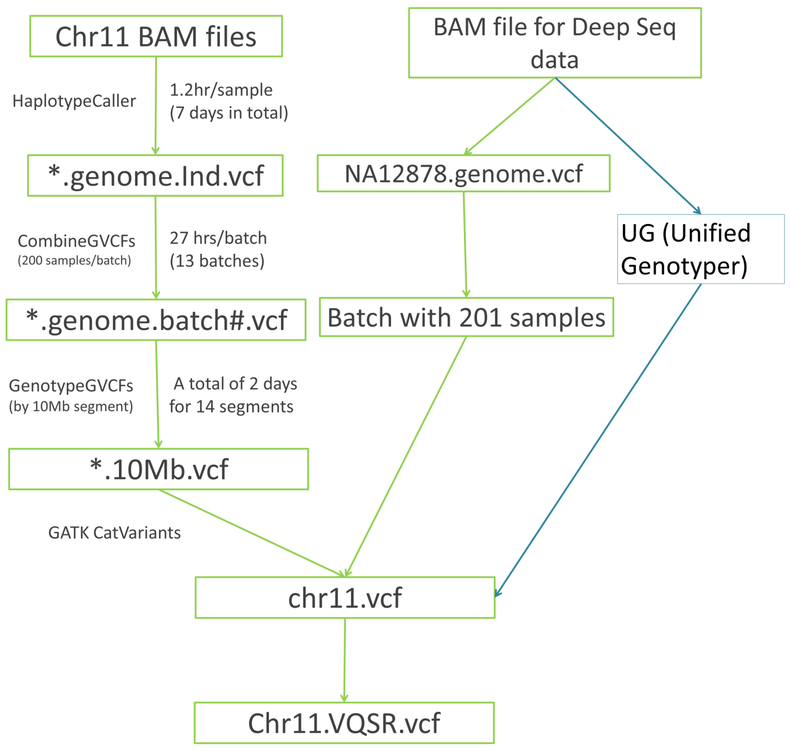
To evaluate the computational feasibility and performance of calling thousands of low pass WGS samples under current hardware and software supports, we used 2,535 low pass WG BAM files from the 1KGP for chr11 alone with 1 CEU trio that were WGS in house by two different library preparation methods. We checked the computational time, evaluated the overall data quality by comparing sequencing data to array data and compared genotype calls made by both UG and HC.

Methods

Input data

- BAM files for chr11 from 1000 genome Phase III (2535 samples)
- CIDR's deep WGS (30x) data for a CEU trio w/Library Prep from
  - Regular
  - PCR-free

Data processing

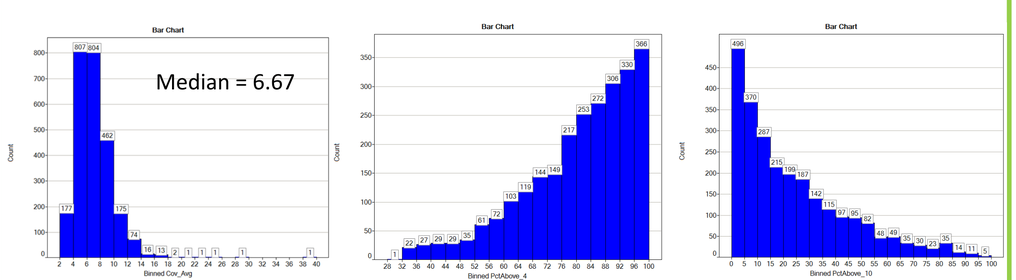


Data Analysis

- Tools: BWA, GATK (3.1-1)
- HetSensitivity = $\frac{(\#GTs = \text{"0/1"} \mid GTg = \text{"AB"}) + (GTs = 1/1 \mid GTg = \text{"AB"})}{(\#GTg = \text{"AB"})}$
- HetConcordance = $\frac{(\#GTs = \text{"0/1"} \mid GTg = \text{"AB"})}{(\#GTs = \text{"1/1"} \mid GTg = \text{"AB"}) + (GTs = 0/1 \mid GTg = \text{"AB"})}$
- HomConcordance = $\frac{(\#GTs = \text{"1/1"} \mid GTg = \text{"AA or BB"})}{(\#GTs = \text{"1/1"} \mid GTg = \text{"AA or BB"}) + (\#GTs = \text{"0/1"} \mid GTg = \text{"AA or BB"})}$

Results

Fig 1: Mean and % Coverage @ 4x and 10x for Low Pass data



Results

Table1: Variant call summary by various calling groups

chr11	Variant Type	2535Callset	NA12878_LowPass_2535Callset	NA12878_LowPass_200callset	NA12878_LowPass_201Callset with Regular LibPrep	NA12878_LowPass_201 Callset with PCRfree	NA12878_DeepRegular_201 Callset	NA12878_PCRfree_201 Callset	GIAB (All PASS)
LowPass_HC(ALL)	ALL	4,110,119	161,367	160,796	161,101	161,111	234,599	228,284	
LowPass_HC(PASS)	ALL	4,059,964	159,334						193,386
LowPass_HC(ALL)	SNV	3,673,655	135,866	136,688	137,524	137,528	188,592	187,343	
LowPass_HC(PASS)	SNV	3,636,687	133,968						178,099
LowPass_HC(ALL)	INDEL	436,464	25,501	24,108	23,577	23,583	46,007	40,941	
LowPass_HC(PASS)	INDEL	423,277	25,366						15,287
LowPass_HC(ALL)	Singleton	1,504,744	308						

- Calling variants in larger batch seems to call a little bit more variants for individual sample.
- Including deep sequencing data for NA12878 has little effect on the total # of variants called by the low pass data.
- For SNVs called by Deep Seq only, they have significantly low DP in LowPass data (2.24 vs 5.25) and about 25% sites have DP of 0 (Data now shown).
- 36% variants are singletons.

Table2: Sensitivity and concordance of sequencing data to array data for SNVs

Metrics	WGS Data	GIAB_v2.18	OmniExpress	Omni1M	Omni2.5
HetSensitivity	NA12878_LowPass	63.87%	67.24%	66.82%	64.38%
HetSensitivity	NA12878_LowPass_PASS	63.62%	67.23%	66.81%	64.35%
HetSensitivity	NA12878_LowPass_200Callset	64.23%	67.40%	66.98%	64.59%
HetSensitivity	NA12878_Deep_Regular	99.17%	99.85%	99.76%	97.05%
HetSensitivity	NA12878_Deep_PCRfree	99.18%	99.84%	99.75%	97.04%
ConcordanceHet	NA12878_LowPass	89.81%	90.21%	90.08%	89.97%
ConcordanceHet	NA12878_LowPass_PASS	89.82%	90.21%	90.08%	89.98%
ConcordanceHet	NA12878_LowPass_200Callset	89.79%	90.19%	90.07%	89.96%
ConcordanceHet	NA12878_Deep_Regular	99.98%	99.96%	99.98%	100.00%
ConcordanceHet	NA12878_Deep_PCRfree	99.99%	99.96%	99.98%	100.00%
ConcordanceHom	NA12878_LowPass	99.92%	99.82%	99.73%	99.56%
ConcordanceHom	NA12878_LowPass_PASS	99.70%	99.82%	99.74%	99.64%
ConcordanceHom	NA12878_LowPass_200Callset	99.81%	99.83%	99.73%	99.55%
ConcordanceHom	NA12878_Deep_Regular	99.72%	99.78%	99.72%	98.60%
ConcordanceHom	NA12878_Deep_PCRfree	99.82%	99.84%	99.76%	99.34%

- Both Heterozygous Sensitivity and Concordance is much higher for deep sequencing data than Low Pass data (65 vs 99% for Het Sen & 90 vs 99.9% for Het Concordance)
- Homozygous concordance is quite similar between low pass and deep sequencing data.
- For Deep Seq data, PCR free LibPrep has a little improvement in HomConcordance than regular LibPrep.

Table 3a: Variant call summary for deep sequencing data called by UG and HC

Deep Seq	SNV	INDEL	Total
Regular_UG	199,687	40,420	240,107
Regular_HC	188,662	45,937	234,599
PCRfree_UG	260,951	39,719	300,670
PCRfree_HC	187,417	40,867	228,284

- HC called more INDELs and less SNVs
- Majority of SNVs missed by HC are of low quality
- INDELs have much lower concordance than SNVs.
- For sites that were mixed of SNV and INDEL, single sample level SNV were grouped into INDEL in HC, resulting in underestimate of SNV count by HC.

Table 3b: Comparison of variant calls between UG and HC

	Regular Library Preparation			PCRfree Library Preparation		
SNV	PASS VQSR in UG	Failed VQSR in UG	Total	PASS VQSR in UG	Failed VQSR in UG	Total
Same	161,940	21,785 (12%)	183,725	161,181	22,232 (12%)	183,413
Diff	119	554 (82%)	673	77	838 (92%)	915
MissByHC	279	15,010 (98%)	15,289	51,917	24,706 (32%)	76,623
MissByUG			4264			3,089
INDEL	PASS VQSR in UG	Failed VQSR in UG	Total	PASS VQSR in UG	Failed VQSR in UG	Total
Same	25,987	5,284 (17%)	31,271	27,513	5,228 (16%)	32,741
Diff	2,989	937 (24%)	3,926	2,001	897 (31%)	2,898
MissByHC	2,409	2,814 (54%)	5,223	1,795	2,285 (56%)	4,080
MissByUG			10,740			5,228

Summary

- Time-wise, it is feasible to use HC to do variant calling for studies with a few hundred samples though generating genomic vcf files are the most time-consuming step.
- Low Pass WGS (NA12878 @ 4.91x) has 70% sensitivity to GIAB, and 90% Het concordance to array data which is much lower than Deep (30x) WGS.
- Besides lower sensitivity due to insufficient coverage, the low Het Concordance rate for Low Pass data suggests the difficulty in differentiating Het from AltHom which may be a major challenge for low pass WGS.
- # of variants per samples seems to increase as batch size increases. However, having deep sequenced sample included in the callset did not seem to improve sensitivity for low pass sample.
- For deep sequencing data, HC seems to call less total variants and SNVs than UG. The vast majority of SNVs missed by HC were of low quality.