Utilizing the Genome Analysis Toolkit's (GATK) CalculateGenotypePosteriors to refine sequencing genotype calls based on external population and trio information

Elizabeth Pugh, Kurt Hetrick, Peng Zhang , Kimberly Doheny

Center for Inherited Disease Research, Institute of Genetic Medicine, The Johns Hopkins School of Medicine, Baltimore, MD

Motivation

GATK 3.3-0 can use population allele frequency data and/or trio information to refine the posterior probability of the genotypes generated by HaplotypeCaller. We hoped to use genotype refinement to reduce the number of false positive de novo variants seen without losing 'real' de novo variants.

Data and Methods

WES 767 whole exome (WES) samples (Agilent® SureSelect™ XT Human All Exon v4, Illumina® HiSeq™ 2000/2500) sequenced at CIDR as part of our work for the Baylor Hopkins Center for Mendelian Genomics (http://www.mendelian.org/) containing a variety of ethnicities and family structures.

WES Variant Calling

BAM files were generated using GATK 3 and bwa mem 0.7.8

gVCF files were created with GATK 3-3.0 HaplotypeCaller on baited regions (79 Mb), joint called with GenotypeGVCF

variant sites filtered with Variant Quality Score Recalibration (VQSR) - Best Practices

WGS Hapmap trio processed with GATK Best Practices on Bina RAVE module using GATK v2014.3-3.2.2-7-gf9cba99 and bwa-mem 0.7.8

Other Methods

WES Genotypes were refined with CalculateGenotypePosteriors (CGP) using different combinations of supporting external population information (1KG phase 3 version 5 and/or Exome Aggregation Consortium Release 0.2) and/or pedigree information.

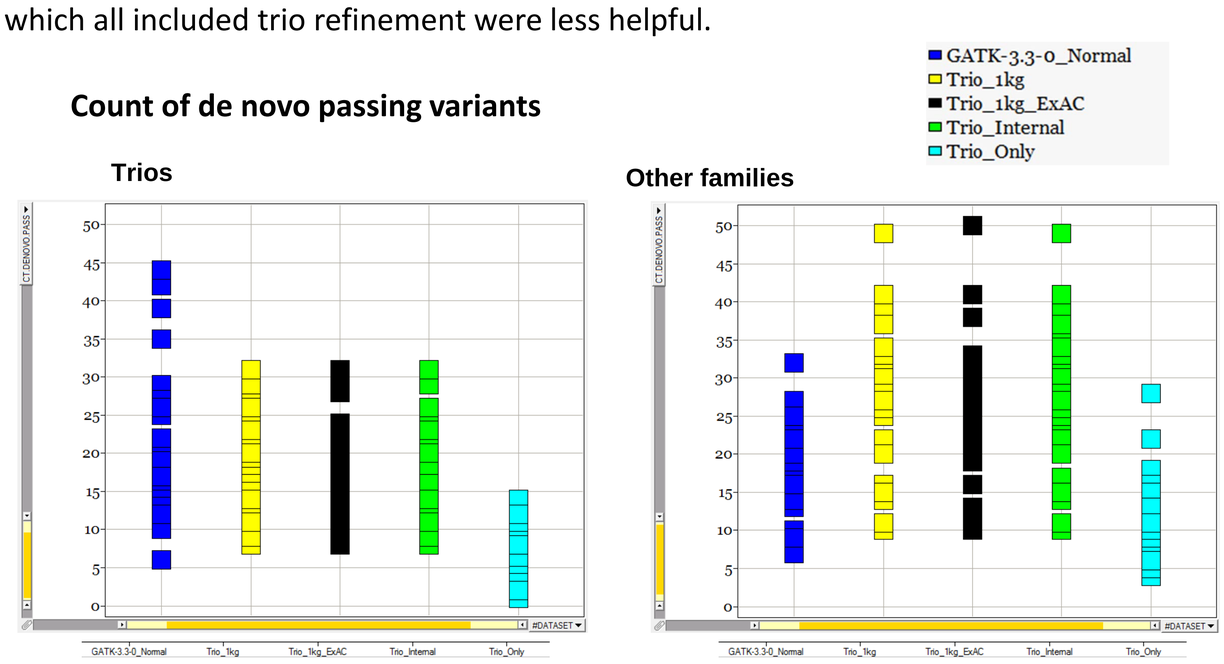
WGS Genotypes were refined using Trio plus internal population frequencies

Concordance and sensitivity to genotypes generated on a SNP array was calculated for each sample.

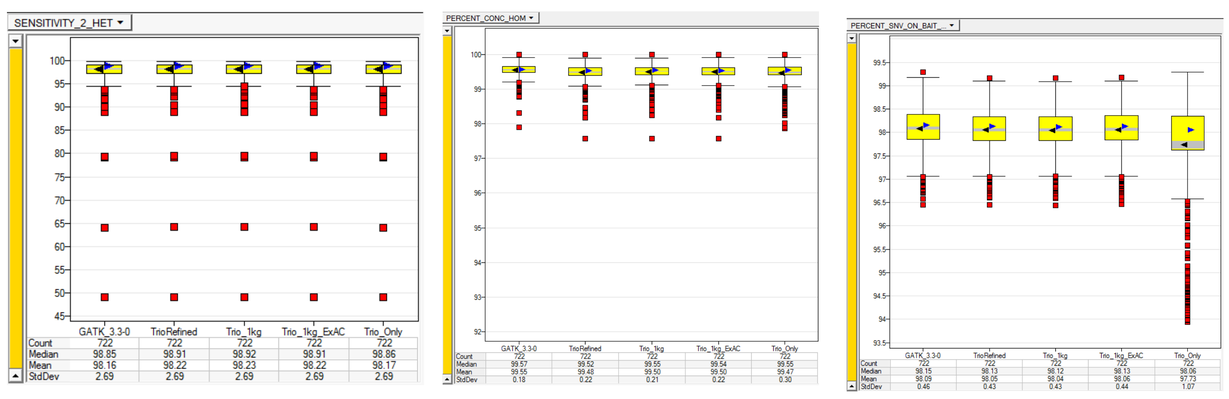
Mendelian errors were calculated using PLINK.

Results

For the trio subset of WES families, the combination of trio and frequency refinement reduced the mean count of de novo passing variants from 20 to 16 while with trio refinement alone it was 5. Unexpectedly, for other family structures these strategies which all included trio refinement were less helpful.



Sequencing quality metrics Mean sensitivity to heterozygotes and homozygote concordance were similar across refinement strategies. The percent of SNPs in dbSNP 138 dropped for a subset of samples when only trio refinement was used with 78 samples having less than than 96.5% of SNVs in dbSNP138.



WGS results

The percent of Mendelian inconsistent calls was reduced for both SNVs and INDELs after trio genotype refinement.

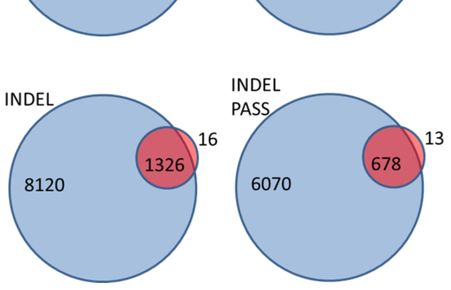
sample	% missing	% inconsistent	% missing	% inconsistent
PASS SNV BEFORE CGP			PASS INDEL BEFORE CGP	
NA12878	**0.25%**	**0.60%**	**1.40%**	**6.32%**
PASS SNV AFTER CGP			PASS INDEL AFTER CGP	
NA12878	**0.21%**	**0.08%**	**1.40%**	**0.15%**

de novo Variant Calls Before and After CGP

Before Refinement
After Refinement

	Before_Refinement	After_Refinement
SNV	12933	4269
SNV_HI	9371	393
SNV_LO	3562	3876
SNV_PASS	6078	2765
SNV_PASS_HI	4788	46
SNV_PASS_LO	1290	2719

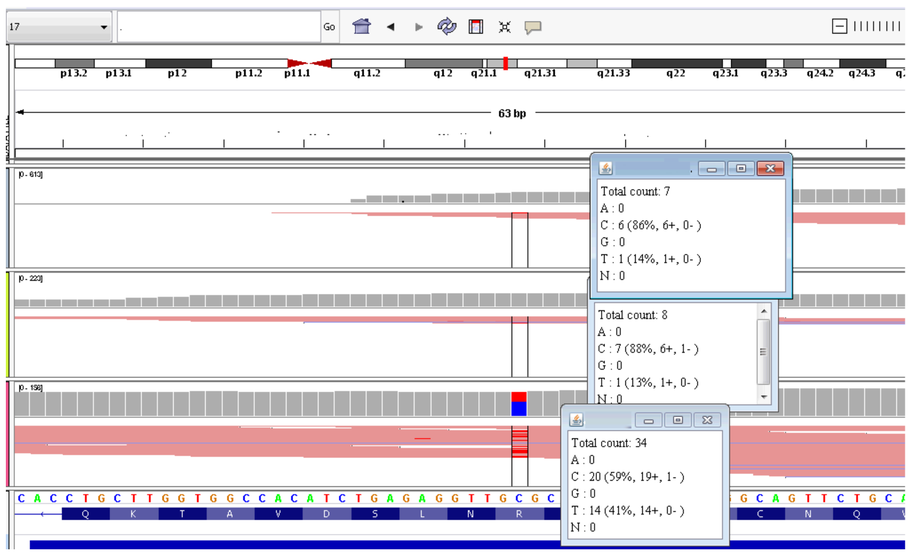
	Before_Refinement	After_Refinement
INDEL	9446	1342
INDEL_HI	5393	153
INDEL_LO	4053	1189
INDEL_PASS	6748	691
INDEL_PASS_HI	3463	34
INDEL_PASS_LO	3285	657



The table and Venn diagrams show the change of the de novo variant calls after refinement. The majority of the de novo calls are no longer made, a subset remain along with a few new calls.

HI LO Confidence DeNovo Definition: "Low confidence DeNovos have child GQ >= 10 and AC < 4 or AF < 0.1%, whichever is more stringent for the number of samples in the dataset. High confidence de novo sites have all trio sample GQs >= 20 with the same AC/AF criterion."

"Real Variants' To evaluate whether genotype refinement overcorrected and we lost 'real variants' we checked 86 unique variants (76 SNVs and 10 indels) that had been previously identified as potentially related to disease and verified by Sanger sequencing. All genotypes remained unchanged except for one SNV site which was previously a singleton was additionally called in two unrelated individuals (from other trios) when trio refinement alone was used.



IGV plots show three unrelated individuals. The bottom sample was called originally. The top two are new calls for just the trio only refinement strategy.

Conclusions

Trio refinement works well on pure trios but may not be as useful for nuclear families and extended pedigrees.

We are currently exploring using population refinement alone for sets of unrelated samples. Preliminary results (not shown) suggest this also works well.

As always there is a trade off between sensitivity and specificity. As such we discuss the use of refinement strategies with the project PI and analysts and have released data with and without refinement for the same project.