Accuracy of whole genome sequencing and the Platypus variant caller in identifying genetic variation within the structurally challenging Epidermal Differentiation Complex

Claire Malley[1], Meher Boorgula[2], Sameer Chavan[2], Nicholas Rafaels[2], Donald Leung[3], Kathleen Barnes[2], and Rasika Mathias[1]

On behalf of the Atopic Dermatitis Research Network

1. Johns Hopkins University, Division of Allergy & Asthma, Baltimore, Maryland, USA.
2. Translational Informatics and Computational Resource (TICR), Colorado Center for Personalized Medicine (CCPM), Denver, CO, USA.
3. Pediatric Allergy and Clinical Immunology, National Jewish Health, Denver, CO, USA.

BACKGROUND

❖ Segmental duplications and tandem repeats in the Epidermal Differentiation Complex (EDC; hg19 chr1:151972910-153642037) are challenging for high-throughput sequencing

❖Platypus variant caller software calls SNPs and indels simultaneously in a joint, multi-sample framework. Illumina Isaac software calls SNPs and indels separately, one sample at a time. Joint calling improves accuracy of the variant type determination and of genotypes, particularly for rare variants.

❖The new joint calling framework is preferable because genotype calls and variant types can be more accurate than those from Illumina Issac

STUDY AIM

Apply the Platypus variant caller to integrate the calling of SNPs and indels within a single framework, and evaluate the accuracy of variant calling within the EDC.

METHODS

❖Three sets of genetic data were available on 799 European American subjects from the ADRN:

1. 30X WGS generated on the Illumina HiSeq
2. OMNI 2.5 GWAS SNP array
3. TaqMan genotypes on a set of four variants in the filaggrin (FLG) gene of prior clinical interest: 2282DEL4, R2447X, R501X, and S3247X.

❖Platypus variant caller was run on all subjects to create multi-sample variant call files (VCF) containing both SNPs and indels. Only sites that passed the variant caller flags and genotypes with coverage > 7 and genotype quality (GQ) > 20 were retained for analysis.

❖The pipeline was run on three additional gene regions (STAT6, IFNG, IL4R) of equal length as the EDC but with fewer segmental duplication or tandem repeat issues for comparison.

❖Platypus SNP calls were compared to the OMNI and TaqMan genotype data to determine concordance per individual and per variant. OMNI reference and homozygous genotypes were systematically checked then corrected for allele strandedness, since Illumina OMNI arrays use TOP/BOTTOM scheme rather than REF/ALT.

CODE REPOSITORY

GitHub
github.com/cemalley/ASHG17
Platypus is available at well.ox.ac.uk/platypus

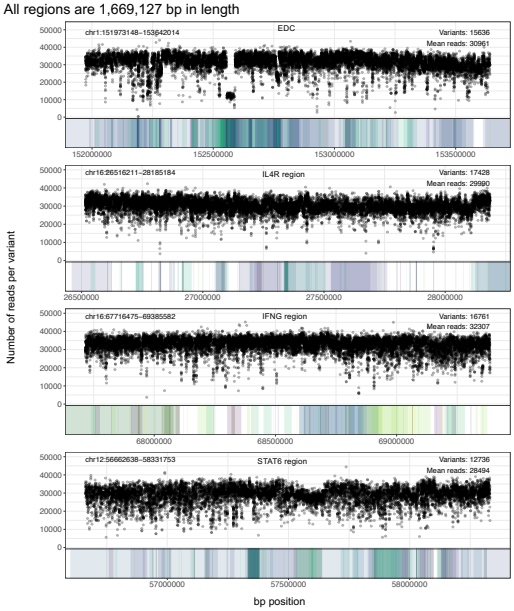
❖Funded by National Institute of Allergy and Infectious Diseases, National Institutes of Health, Department of Health and Human Services, under contract no: U19AI117673

RESULTS

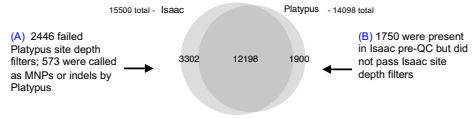
Platypus coverage for EDC is high despite SVs

In sections of lower coverage for the EDC, we found greater segmental duplications and CNVs. Despite denser structural variants in the EDC, mean coverage and variant counts were comparable to the three other regions of equal size, where there were fewer structural variants.

All regions are 1,669,127 bp in length



Bi-allelic SNPs in EDC: Platypus vs. Illumina Isaac variant calls



Although the data input for Platypus and Isaac were the same Illumina HiSeq bam files, there were SNP calls exclusive to both due to differences in the softwares' quality filtering algorithms:

(A) Platypus called 2446 variants with the non-passing 'QD' site filter flag. At least 573 were indels and multi-nucleotide polymorphisms where Illumina called passing SNPs.
(B) Illumina Isaac site and genotype filters removed 1750 variants which did pass Platypus filters; 'LowGQX' predominates in the single-sample filter flags.

Platypus variant calls vs. Illumina OMNI 2.5 SNP array

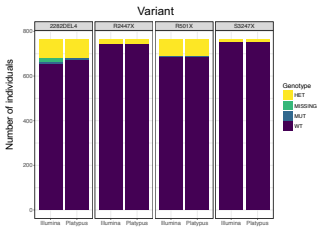
99.9% mean concordance (range 98.5-100) for a subset of 850 non-monomorphic transition overlapping SNPs called in both OMNI and WGS data.

Concordance for filaggrin (FLG) variants

N total individuals in common = 766

Platypus vs. TaqMan
Where N = non-missing genotypes:
2282DEL4: N = 749, 11 discrepancies or 98.5% concordant
R2447X: N = 638, 1 discrepancy or 99.8% concordant
R501X: N = 735, 100% concordant
S3247X: N = 668, 100% concordant
Discordant TaqMan genotype calls were manually inspected and the Platypus calls were laboratory-confirmed to be correct.

Platypus vs. Illumina Isaac
Platypus and Illumina Isaac were 100% concordant for non-missing genotypes, but Isaac had higher rate of missing calls for the 2282DEL4 indel (20/766 or 3%). Platypus determined the individuals were WT.



Genotype codes: HET = heterozygous, **MISSING** = individual's genotype could not be called; **MUT** = homozygous for mutation or alternate allele; **WT** = wild type, as in homozygous for reference allele.

Summary of Results

There were 15,636 SNPs passing QC filters in the EDC with Platypus with mean total coverage per variant of 30,971 (range: 337-46,696). Overall coverage and variant calling statistics were comparable to the three other gene regions (STAT6, IFNG, IL4R). We found a mean of 99% (range: 98-100%) genotype concordance between Platypus and OMNI. Additionally, there was 98-100% genotype concordance with TaqMan called FLG variants.

CONCLUSIONS/IMPLICATIONS

❖The Platypus variant caller provides high quality variant calls from whole genome sequencing data rife with structural variation, which has led to unreliable or impossible calls in the past.

❖Platypus-called genotypes are highly concordant with older-generation genotyping array and qPCR data, occasionally more accurate than TaqMan.

❖We are confident in the quality of the software's joint calling for generating squared-off multi-sample VCFs of both SNPs and indels.

❖Platypus may be the most effective caller for the EDC and similar structurally complex regions, particularly for indels, compared to Illumina Isaac