Assessing SNP using the KPGP-38 Human Genomes next-generation sequencing data from CAMDA

Valerii Soika[1], Wenqian Zhang[2], Jie Shen[2], Joe Meehan[2], Zhenqiang Su[2], Weigong Ge[2], Hong Fang[3], Roger Perkins[2], Huixiao Hong[2], Weida Tong[2], Vahan Simonyan[1]

[1]*Office of The Center Director, Center for Biologics Evaluation and Research, U.S. Food and Drug Administration, Rockville, MD 20892, USA*
[2]*Division of Bioinformatics and Biostatistics, National Center for Toxicological Research, U.S. Food and Drug Administration, Jefferson, AR 72079, USA*
[3]*Office of Scientific Coordination, National Center for Toxicological Research, U.S. Food and Drug Administration, Jefferson, AR 72079, USA*

ABSTRACT

The Critical Assessment of Massive Data Analysis (CAMDA) consortium hosts the Next Generation Sequencing (NGS) data of 38 human genomes from the Korean Personal Genome Project (KPGP-38). The high sequencing coverage and the inclusion of two pairs of twins make KPGP-38 a suitable data set to explore quality control metrics for improving accuracy in SNP and genotype calling and to evaluate performance of NGS data analysis tools.

We used the FDA's High-performance Integrated Virtual Environment (HIVE), a cloud-based environment optimized for the storage and analysis of extra-large data (primarily NGS data), to align the NGS data of the two pairs of twins in KPGP-38 to the reference human genome. HIVE was used to align raw reads to the reference genome, to call and compare SNPs. SNP concordances between the four subjects were further analyzed to evaluate HIVE's performance. Our results revealed that SNPs from twins are in high concordance (~94%) while non-twin subjects share fewer SNPs (~69%), indicating HIVE is a reliable tool for NGS data analysis.

WORKFLOW

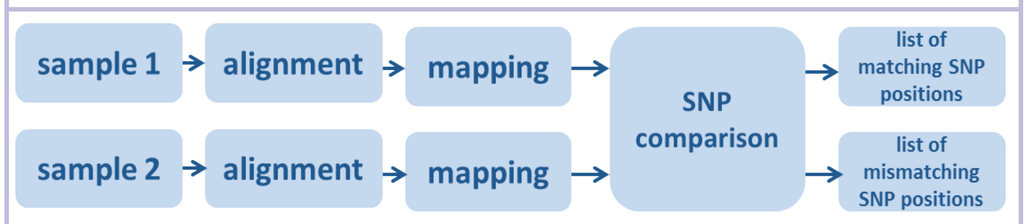


FIGURE 1. HIVE workflow for SNP detections and comparison

Sequencing data for two homozygous twins are considered.
Each sequencing data set is about 225 GB in size.
Data for each sample are divided into 14 files, 28 files for 2 samples, about 450 GB total.

Parameter	Sample 1	Sample 2
Reference	A	A
Consensus	C	G
Count-A	0	2
Count-C	35	21
Count-G	27	23
Count-T	2	7
Count-Insertions	0	0
Count-Deletions	0	0
Count Total	64	53
Count Forward	49	36
Count Reverse	15	17
Quality	47	46
Entropy	0	0
SNP Entropy	0	0
Freq A	0	0
Freq C	55	40
Freq G	42	43
Freq T	3.1	13

FIGURE 2. SNP representation in HIVE for a single position

APPROACH

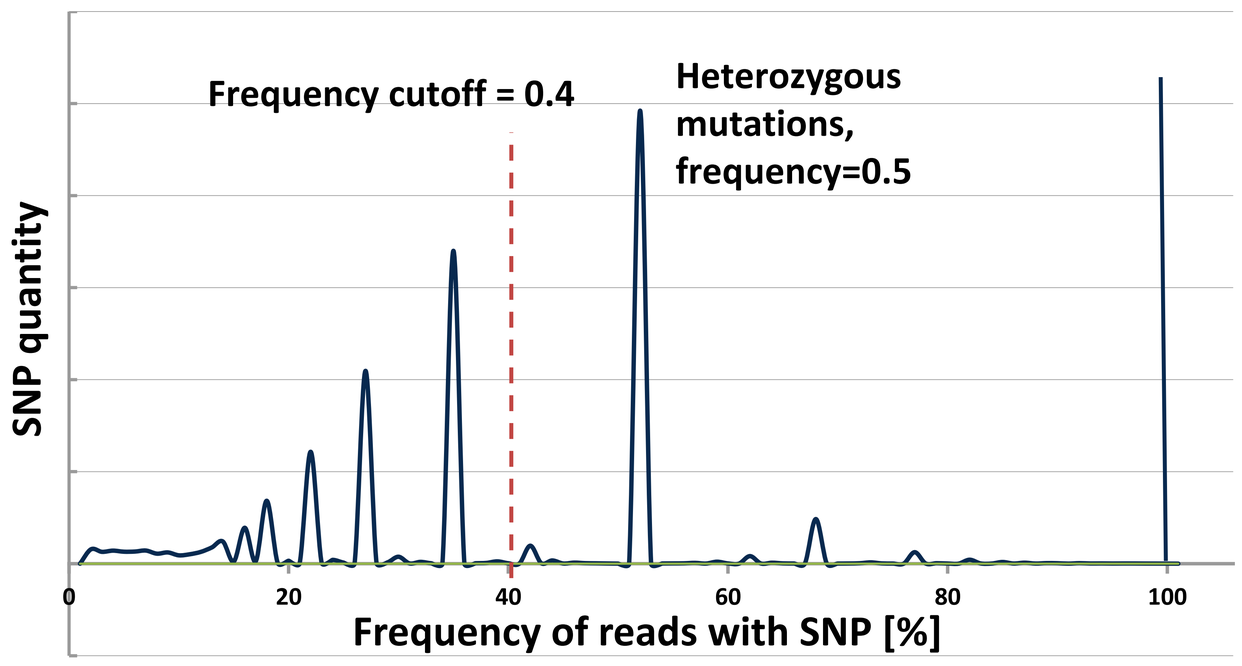
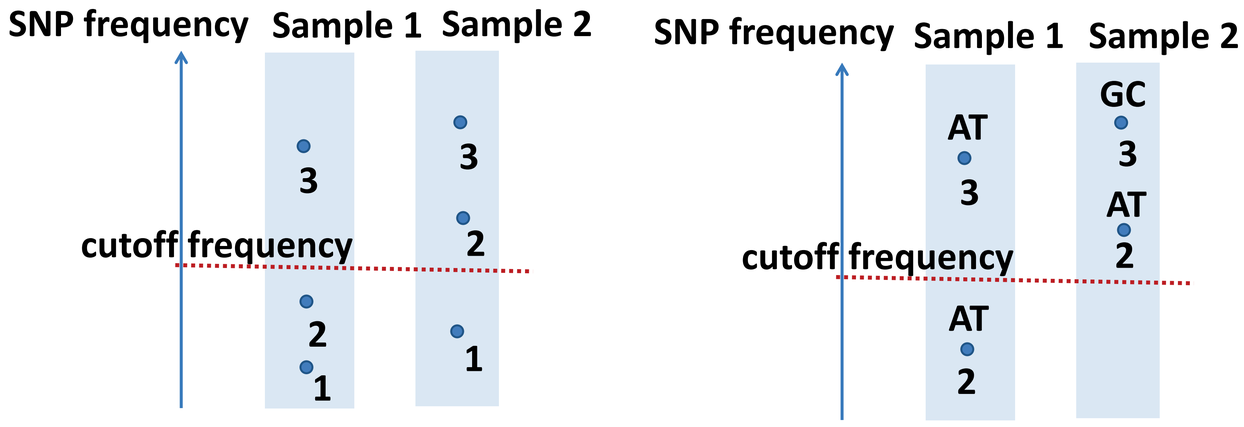


FIGURE 3. Frequency distribution of reads with SNP (SNP frequency) for chromosome 1



Frequency SNP characteristics at a given position
Case 1: both samples have SNP with frequency below threshold
Case 2: only one sample have SNP frequency above threshold
Case 3: Both samples have SNP above threshold
Case 1 is excluded from consideration.

SNP comparison for diploid genome
Case 2: genotype match
Case 3: genotype mismatch

FIGURE 4. SNP comparison approach

RESULTS

samples	matches count	mismatches count	%mismatches
Twins 1 vs 2	4222471	257835	5.8
Twins 3 vs 4	4259950	249053	5.5
Not twins 1 vs 3	3316734	1508904	31.3
Not twins 1 vs 4	3137507	1425283	31.2
Not twins 2 vs 3	3411040	1539265	31.1
Not twins 2 vs 4	3411738	1538355	31.1

FIGURE 5. Detected SNP mismatches for samples from twins and unrelated persons.

Approach detects high SNP similarity for twin genomes.
Lower SNP similarity is detected for the samples from unrelated people.

ACKNOWLEDGEMENTS

Carolyn A. Wilson, PhD, Associate Director for Research, Office of the Center Director, FDA CBER
Konstantin Chumakov, PhD, Associate Director for Research, Office of Vaccines Research and Review, FDA CBER
Implementation: blueHIVE technology group bluehivescience@gmail.com (Mazumder and Simonyan research groups)