Novel HIVE methods in cancer genomics: Mutation detection, phylogentic analysis of patients, scanning against a comprehensive curated cancer database BioMuta

John Torcivia[1], Haichen Zhang[2], Yang Pan[2], Krista Smith[2], Hayley Dingerdissen[1], Vahan Simonyan[1], Raja Mazumder[2]

[1]Food and Drug Administration, Center for Biologics Evaluation and Research, Rockville, MD, 20852

[2]Department of Biochemistry and Molecular Biology, The George Washington University, Washington, DC, 20037

School of Medicine & Health Sciences
THE GEORGE WASHINGTON UNIVERSITY

FDA

ABSTRACT

HIVE, the High-performance Integrated Virtual Environment, is a cloud cluster environment specifically developed to overcome challenges associated with extra-large next-generation sequencing (NGS) data. Through a number of algorithmic advances (including but not limited to exploitation of sequence nature, parallelization, sequence sorting, self-similarity recognition and non-redundification of read sets) the HIVE Cancer Annotation Pipeline is capable of mutation discovery with increased sensitivity requiring less time than previously described methods.

The reference study surveyed tumor and adjacent non-tumor liver genomes extracted from 81 HBV-positive and 7 HBV-negative hepatocellular carcinoma (HCC) samples. Reads were uploaded directly from SRA and analyzed using the HIVE-hexagon aligner and HIVE-heptagon profiler. Preliminary findings show the HIVE approach detects the same mutations reported by the previous reference analysis in addition to several new mutations not previously reported.

BACKGROUND

Previous Study: Sung et. al., 2012

- Sequenced 81 HBV-positive and 7 HBV negative tumor (HCC) and adjacent normal liver tissue
- Mapped paired-end reads to human reference genome (hg19) and HBV genome (NC_003977).
- Gene expression microarray analysis
- Analysis of HBV integration and CNVs (CNVs were found to increase at HBV breakpoint locations with chromosomal instability)
- Hepatitus B integration (HBV) was more frequent in tumor versus adjacent liver tissues (86% vs 30%)

Current Study:

- Mapped paired-end reads to human reference genome (hg19), SNV impact analysis, and phylogenetic analysis

DATA

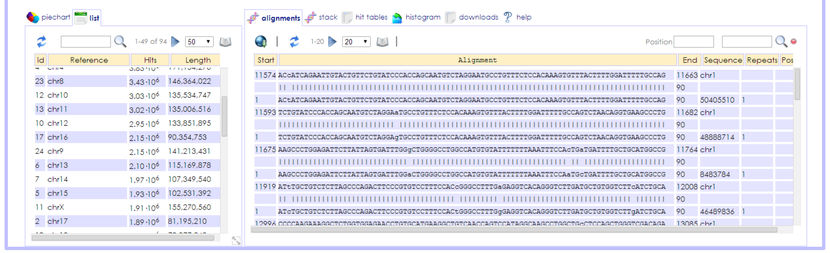
Reads: 15 patients (tumor and non-tumor) paired-end read DNA-Seq data from Sung et. al 2012, obtained from SRA/NCBI

Genome: Hg 19, GRCh37 Genome Reference Consortium Human 37 (GCF_000001405.22) downloaded from NCBI

METHODS

❖ ALIGNMENT

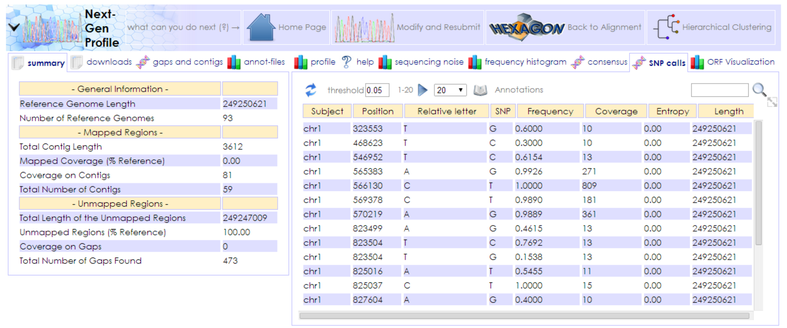
Paired-end reads from patient samples (tumor and normal) were aligned to the reference human genome (Hg19) utilizing the HIVE-hexagon alignment tool.



METHODS (continued)

❖ Sequence Profiling

Aligned reads have been used to call nucleotide base variants along the profile of the reference sequence. Together with the conventional "pileup" approach, base-calling, genotyping and gene expression, the HIVE SNP–profiler additionally calculates frequency maps, histograms, Shannon's entropic information content factors, forward/reverse disbalance, insertions and deletions.



❖ dbSNP Concordance

The HIVE-heptagon SNP profiling results for patient samples can compliment and extend variation information present in dbSNP. The SNP summary was downloaded from the HIVE interface in vcf format to display the detected SNVs that met our threshold. By uploading this file to SeattleSeq Annotation 137 web service, we detected the number of variants that were covered in dbSNP.

❖ Phylogenetic Analysis

Comparative analysis of SNV profiles allows us to better classify the patients and compare variations across the samples. PhyloSNP, our novel tree-building application, takes in output SNV variation data from the HIVE-heptagon profiler and produces phylogenetic trees.

RESULTS

❖ dbSNP Concordance

Table 1 displays results of the dbSNP concordance for the randomly chosen 15 patients. The table columns display the Patient ID, Type of tissue (non-tumor vs tumor), SRA Accession number, number of SNVs found in dbSNP, number of SNVs found with HIVE-heptagon, and the percentage of SNV overlap between dbSNP and HIVE heptagon results.

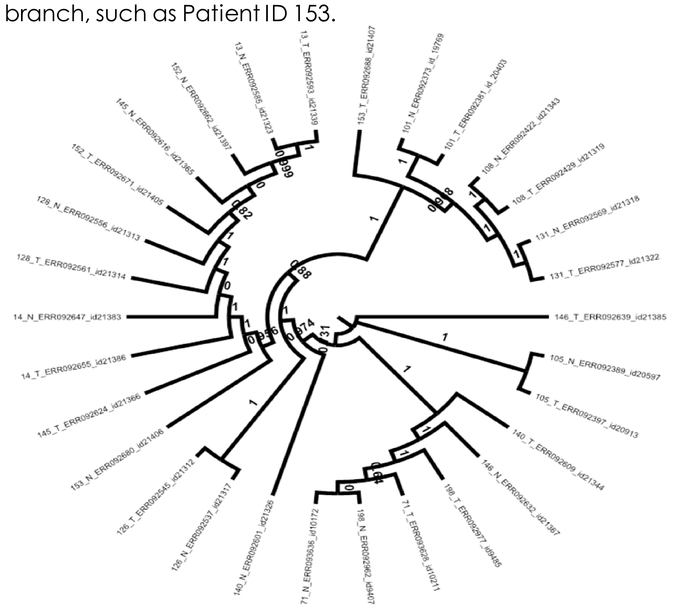
Table 1. dbSNP Concordance

#	Patient_ID	Type	ID	in_dbSNP	Total	Percentage
1	13	N	ERR092585	110982	136423	81.35%
	13	T	ERR092593	96757	119774	80.78%
2	14	N	ERR092647	61498	82965	74.13%
	14	T	ERR092655	27779	38778	71.64%
3	71	N	ERR093628	17740	25030	70.87%
	71	N	ERR093636	10124	13967	72.49%
4	101	N	ERR092373	21012	29090	72.23%
	101	T	ERR092381	18508	25905	71.45%
5	105	N	ERR092389	18468	26390	69.98%
	105	T	ERR092397	22061	31039	71.08%
6	108	N	ERR092422	28303	40922	69.16%
	108	T	ERR092429	21356	30211	70.69%
7	126	N	ERR092573	16147	22724	71.06%
	126	T	ERR092545	24401	33434	72.98%
8	128	N	ERR092556	48640	64535	75.37%
	128	T	ERR092561	48292	63804	75.69%
9	131	N	ERR092569	18812	26424	71.19%
	131	T	ERR092577	16970	23672	71.69%
10	140	N	ERR092601	21549	30529	70.59%
	140	T	ERR092609	17468	25366	68.86%
11	145	N	ERR092616	62887	81217	77.43%
	145	T	ERR092624	39759	52872	75.20%
12	146	N	ERR092632	14893	21647	68.80%
	146	T	ERR092639	19776	28299	69.88%
13	152	N	ERR092662	59949	77688	77.17%
	152	T	ERR092671	57150	74568	76.64%
14	153	N	ERR092680	22577	31390	71.92%
	153	T	ERR092688	16748	23701	70.66%
15	198	N	ERR092962	14003	20091	69.70%
	198	T	ERR092977	19931	27927	71.37%

RESULTS (continued)

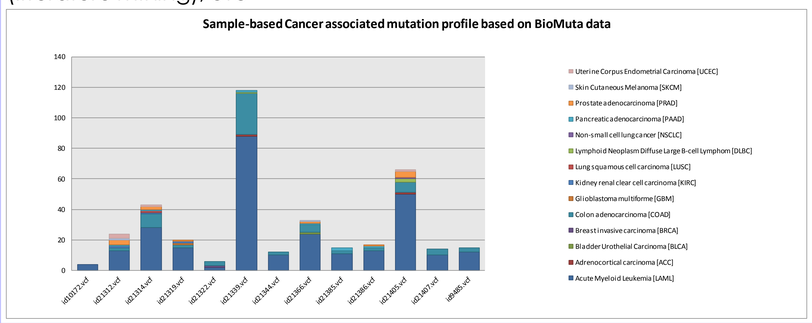
❖ Phylogenetic Analysis

SNV profiles of the patient samples are compared based on the results of the phylogenetic analysis, which determines correlation between case (tumor) and control (non-tumor) samples. It is expected that the normal tissue and tumor tissues from the same patients be located on the same branch. Further analysis is required to investigate the few patient samples (tumor and non-tumor) that are not located on the same branch, such as Patient ID 153.



❖ SNV profiles mapped to Biomuta

We mapped SNVs from each tumor sample to our curated cancer centric database BioMuta, which provides a comprehensive linking from genomic mutation to cancer types. Each individual sample contains a distinct number of SNVs that can be mapped to BioMuta cancer sites and a different pattern of cancer types associated BioMuta curated SNV. Each sample has around 20 of SNVs out of a total average 30K SNVs marked by their overlapping with BioMuta, which integrates all non-synonymous single nucleotide variations from COSMIC, ClinVar, ICGC, publications (literature mining), etc



REFERENCES

Sung, Wing-Kin et. al. **Genome-wide survey of recurrent HBV integration in hepatocellular carcinoma.** *Nature Genetics* **44**, 765–769 (2012)

Presented at the Third Symposium on Translational Genomics Conference. 3/20/14. Bethesda, MD. USA.