Initial characterization of the blood DNA virome from 1000+ Brazilian elderly individuals

Suzana Andreoli Marques Ezquina[1], Michel Naslavsky[2], Maria Rita Passos-Bueno[3], Mayana Zatz[3],

1 Centro de estudos do Genoma Humano - CEGH -USP
2 Centro de Estudos do Genoma Humano - CEGH - USP
3 Centro de Estudos do Genoma Humano - CEGH -USP

Abstract

The characterization of blood DNA virome is important for the identification of emerging pathogens in a population, and an interesting aspect of the integrated virus from infections and immunizations a person has in its lifetime, especially when we are studying people aging 60 years old and beyond. Humans harbor a huge number of endogenous retroviruses embedded in their genome, as remnants of an ancestral germline infection. These endogenous retroviruses may still contribute to pathological processes, including cancer. Whole genome sequencing (WGS) projects that involve whole blood DNA extraction allow the precipitation of external viruses along with the nuclear DNA, which are sequenced and not mapped to the human genome. These viruses contribute to the findings of current infections. Here we intend to present the initial results of the virome analysis, in particular, the determination of viral prevalence and distribution by sex and age. This cohort of 1324 individuals aging from 58 to 104 years (mean age74) is composed by a population-representative sample of São Paulo (SABE study, n=1199) and by a cognitively healthy octogenarians sample (n=125). SABE study individuals harbor the expected incidence of comorbidities under their age range. WGS was carried on the Illumina HiSeqX sequencer using 150 base paired-end single index reads, which were aligned with ISIS Analysis Software to the Human genome version hg38. Duplicate reads were removed. Samtools version 1.5 was used to extract the unmapped reads and the bam file was converted to fastq using bamToFastq and then to fasta using fastq_to_fasta. Blastn+ version 2.6.0 was used to find hits of the fasta files extracted from each individual with the NCBI database "RefSeq and Neighbor nucleotide records" with 116,503 entries of viral genomes. The Blast results were then parsed and counted for each type of virus in each individual using Perl scripts. We filtered putative viral matches from blastn+ using an e-value equal or less than 1e-10. The identification of viruses was performed in raw reads and in contigs after the assembly. In both cases the correlation between the number of reads/contigs and the hits on blast were comparable, but we noted that the assembly of reads in contigs using SOAPdenovo2 yielded overall better hits. Assembled contigs increased the number of longer matches and higher e-values, also decreasing the amount of time needed for the blast program to run. We obtained a mean of 8k hits in 350k queries. In this preliminary analysis, we found that abundances were very similar between women and men, although Human endogenous retrovirus and a strain of Human immunodeficiency virus 2 were slightly increased in women in contrast to Human herpesvirus and mammarenavirus which were slightly increased in men. Further investigation is necessary to better characterize the virome profile of Brazilians in comparison to different populations.

Funding: FUSP