Viral Beacon Use Case: Discovering Shared Intrahost variants

Contents

1 Introduction

In RNA virus infections, intrahost variation emerges from error-prone replication, ending up to multiple circulating quasispecies of low or higher frequency. These variants, in combination with the genetic profile of the host, can potentially influence the natural history of the infection (immunogenicity), the viral phenotype (pathogenicity, tropism), the sensitivity of molecular and serological diagnostics assays and the effectiveness of antiviral drugs and vaccines design.

The exploration of intrahost variability may present also an opportunity to assess viral evolution and viral pathogenicity: Are there some intrahost variants generated specifically in some hosts? Are there some intrahost vriants that are involved in transmission while others are needed/favorable for viral replication within host?

2 Intrahost variants

Viral intrahost diversity is much higher than the interhost viral diversity segregating in the consensus data (GISAID and ENA).

The number of consensus variants in datafreeze 20202405 (combining GISAID & ENA sources) is 26512, while the umber of intrahost variants in (Galaxy LoFreq) is 46359 (33545 (72.3592%) of which are novel, i.e, not found among consensus variants).

Figure 1 shows the greater diversity of intrahost variants vs consensus samples, which is further supported by the presence of much fewer invariant sites (3540 vs 15919) and more polymorphic sites (13412 vs 7362) in intrahost variation data vs consensus data.

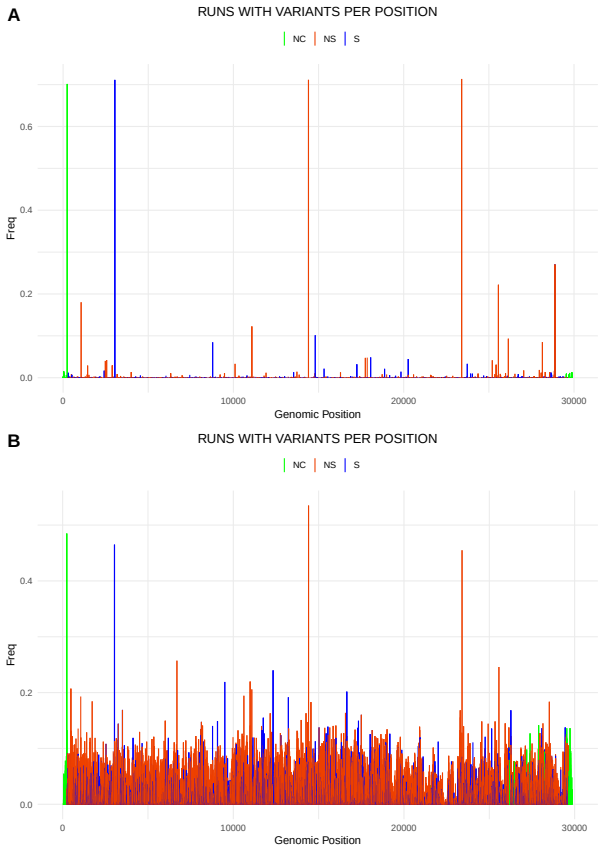

Figure 1: A) Frequency in dataset (matching runs) per genomic position in consensus variants (GISAID). B) Frequency in dataset (matching runs) per genomic position in intrahost variants (Galaxy LoFreq)

Intrahost variants are variants in polymorphic sites that are present in a minor subpopulation of the viral quasiespecies. Although some of these may represent deleterious variants carried on non-replicative genomes, it has been proven in other viruses that a fraction of these could be real viral variation occurring intrahost (Renzette, 2017).

Evidence supporting real variation include the non-random distribution of polymorphic sites in intrahost variants and the presence of shared variants, including aminoacid changing variants among them.

2.1 Minor intrahost variants

In order to focus on minor (subconsensus) intrahost variants, Illumina LoFreq variants also found in consensus datasets (GISAID or ENA) or found at AF above 0.4%AF were removed from further analysis.

A total of 33161 minor subconsensus variants where found.

The relation between AF and matching samples of the 33161 minor subconsensus intrahost variants (figure 2) shows private or rare variants, as well variants shared by a relatively large number of samples.

Interestingly, there are low AF subconsensus intrahost variants that are present in a relatively high number of samples (>10%) (figure). Those are probably the result of convergent evolution/homoplasies and thus could represent epidemiologically/pathologically-relevant intrahost variants that are generated within hosts.

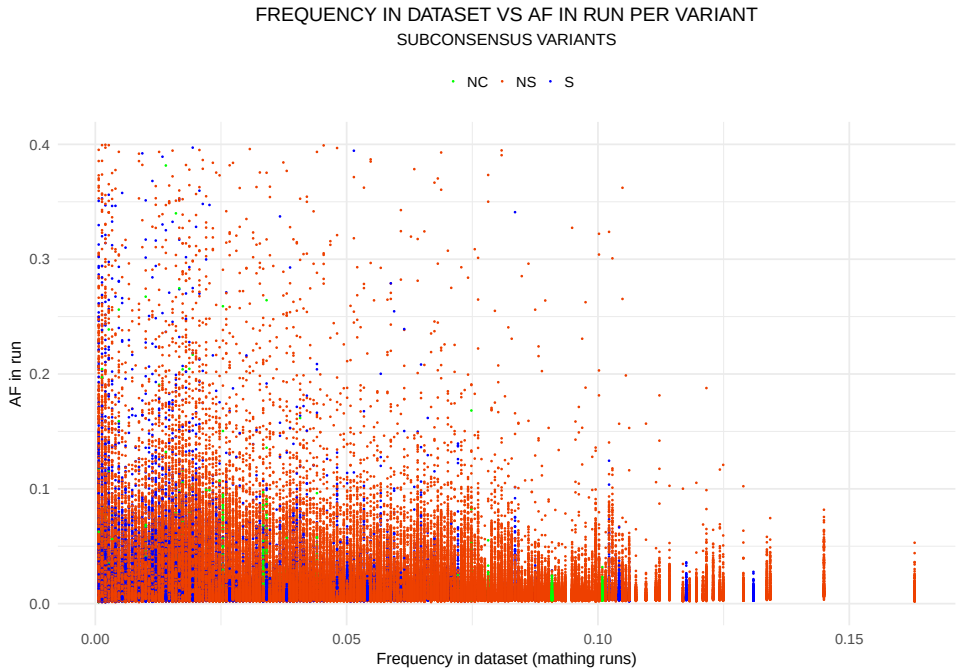

Figure 2: Relation of AF in run vs Frequency in dataset (runs with variant)

Minor intrahost variants are present in all 11 genes and also in NC regions, with a correlation coeficient of 0.9929579 between number of variants and protein-coding region length (Figure 3 panel A), showing a seemingly random distribution of variants.

2.1.1 Shared minor intrahost variants

48 minor intrahost variants were found to be shared among 10% of samples or more (total unique 523 samples)

Interestingly, most of these are NS and AF of NS is also higher in this group than AF of S (p=2.2e-16, Welch Two sample T test)

Shared intrahost variants landed on a subset of genes, namely, "orf1ab", "S", "ORF3a", "N" and "ORF10", with a correlation coeficient of 0.6468671 between number of variants and protein-coding region length (Figure 3 panel B), showing a non-random distribution of variants. Also, no shared intrahost variant landed on NC regions.

It is interesting the prevalence of aminoacid-changing intrahost variants in many replicase components, including the the RNA-dependent RNA polymerase (6 aa substitutions), the proofreading exonuclease 3'-5'ExoN (5 aa substitution), the endoRNAse (3), the helicase (5), and the nsp10/2-O-MT complex (1/1 aa substitutions), as well as in the main protease 3C-like (6 aa substitutions).

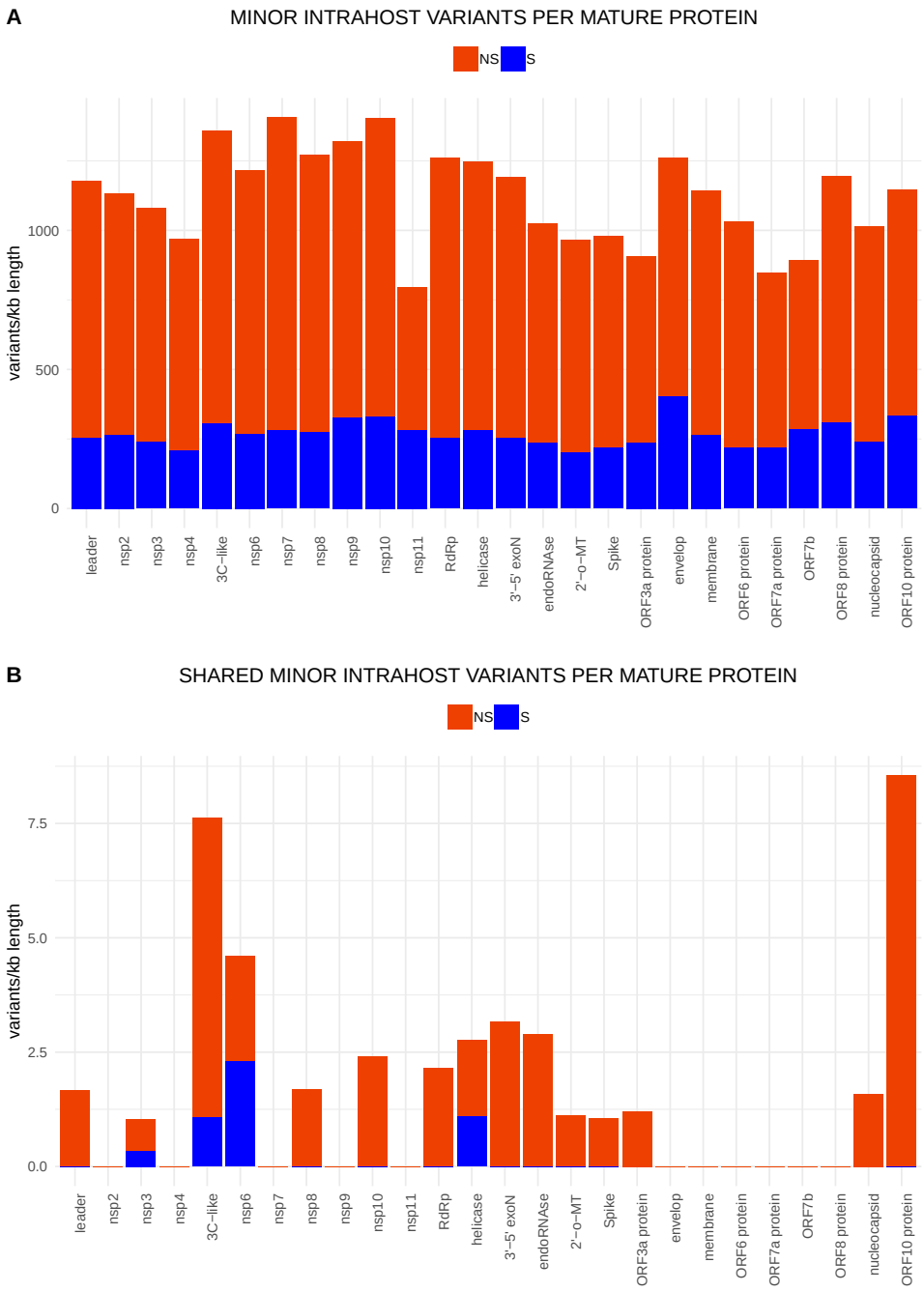

Figure 3: A) Distribution of minor intrahost variants in mature protein coding regions. B) Distribution of shared (>10% samples) minor intrahost variants in mature protein coding regions

Likewise, the prevalence of minor intrahost variants in structural proteins nucleocapsid and Spike, giving 2 and 4 distinct aminoacid changes, respectively, is an interesting finding that merits further

Interestingly, the highest density of aa changing variants landed on ORF10 protein, which is presumably not expressed and which has been associated to the high contagiousness of this virus.

Further investigation these positions.. might be important in confering tissue tropism or some other physiologically interesting phenotype?

Current datafreeze is geographically biased with most samples (and more than half minor variants) coming from Australia. It would be interesting to see if this holds upon broader sampling and see whether there might be an association with geographical origin or phylogenetically related samples.