# Viral Beacon Use Case: Finding Invariant Positions, Monomorphic positions, Polymorphic sites and Intrahost variants

## Contents

# 1 Introduction

Intra host variability of pathogenic viruses and bacteria represents a significant barrier in the control of infectious diseases.

In RNA virus infections, intrahost variation emerges from error-prone replication, ending up to multiple circulating quasispecies of low or higher frequency that could . These variants, in combination with the genetic profile of the host, can potentially influence the natural history of the infection, the viral phenotype, the sensitivity of molecular and serological diagnostics assays and, importantly, it poses a challenge for 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 for viral replication within host?

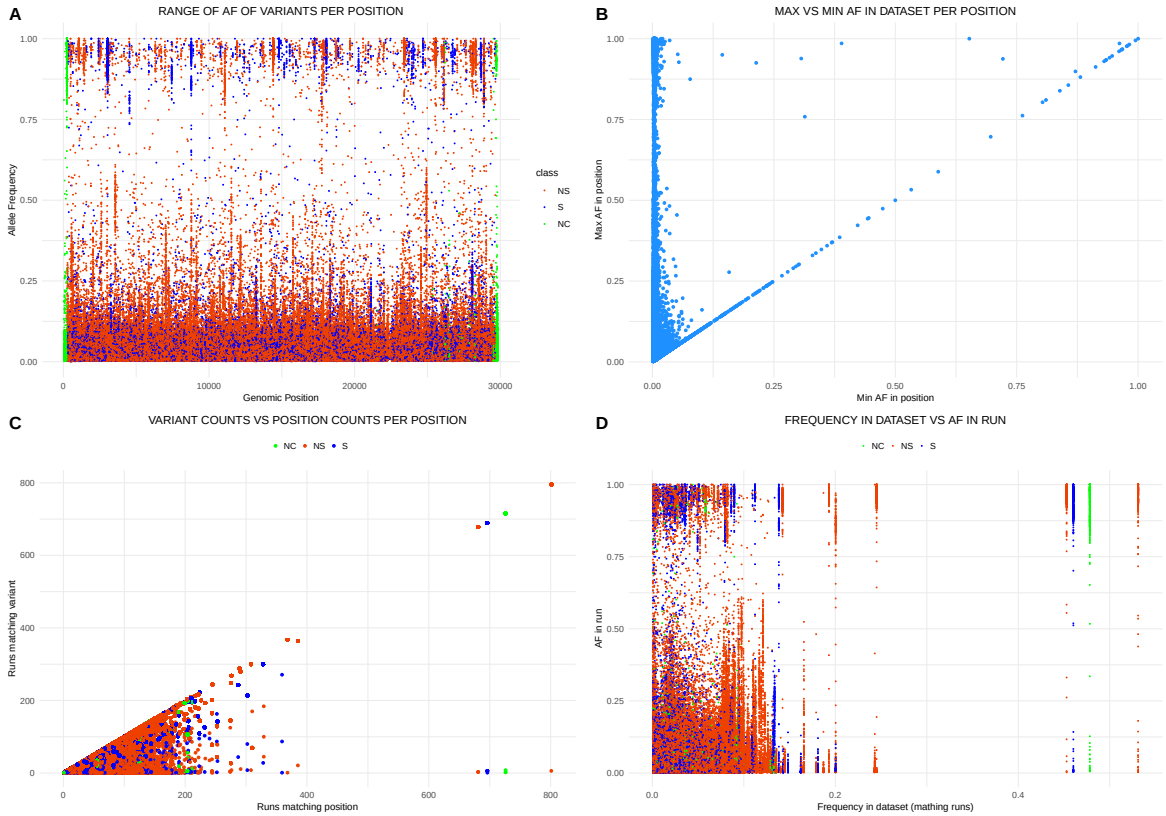


Figure 1: A) Range of AF in dataset of variants at each genomic position. Long lines imply that positions have high frequency and low frequency variants in dataset. B) Relation between max and min AF in dataset per position, variants outside the rect represent variants in polymorphic sites. C) Relation between the number of runs bearing a variant and their maximum AF in dataset. Variants under the the rect represent variants at polymorphic sites, hitting more often a small fraction of runs. D) Relation between the AF in run and the frequency in dataset (proportion of runs bearing the variant), showing variants with relatively low AF that are present in a high number of runs

.

# 2 Inventory: Invariant, Non-Polymorphic and Polymorphic sites

Using SRA Illumina - Galaxy data we can determine polymorphic sites and intrahost quasiespecies.

There are positions showing no variants so far, while others show a unique alternate, sometimes in complete discordance (¿.98 % dataset frequency and AF) and many positions show polymorphism at dataset level and sample level.

Whether these numbers are compatible with neutral evolution or are indication of positive or negative

selection need to be explored.

## 2.1 Invariant positions

3540 (11.8%) genomic positions show no variants in any run in the set, and also not in consensus sequences. Are these positions enriched in some function or some genomic region?
Also, some positions show very little variation. One such contiguous stretches is found in ...see.

## 2.2 Monomorphic

11135 (37.2%) positions show only one alternate.
Some of them are even very pervasive. Is this showing a negative selection?

## 2.3 Positions in absolute discordance

244 positions show one fixed alternate in at least one sample (AF ¿0.98%) How many of these major variants are prevalent in population? 10/244 are mutated in 0.1% of runs, 6/10 in 0.2%, 4 in 0.4% and 1/244 in ¿0.5% How many are less prevalent or private?
Have to check runs coming from same sample. 7/244 are present in 4/1497 runs 10/244 are present in 3/1497 runs 11/244 are present in 2/1497 runs 14/244 are present in just 1/1497 run

## 2.4 Positions with reversions

Variation is loss over time -¿ Pending, using collection date

## 2.5 Polymorphic sites

15231 (50.9%) positions present more than 1 variant (alternate) (2: 10469 (35.0%) 3: 4762 (15.9%))

- Number of positions that are polymorphic at sample (run) level
- Number of runs with variants at polymorphic positions
- Number of shared polymorphic positions (polymorphic at sample level in ¿ 1 sample (run))

## 2.6 Intrahost variants

Intrahost variants are variants in polymorphic sites that are present in a minor subpopulation of the viral quasiespecies. Interestingly, there are variants that have a low frequency in population (dataset), but relatively many samples bear it (figure ) panel D, which should represent intrahost variants.

Some intrahost variants that are never major ones, (never found above 0.4%AF or in consensus data (GISAID).

### 2.6.1 Isolate-specific variants

Some intrahost variants are exclusive to one sample. Variants found in just one sample¿ This needs checking of runs from same sample

Some intrahost variants are common to many samples. Are the latter a signal of convergent evolution/homoplasies?

### 2.6.2 Intrahost variants molecular consequence class

% of correspond to missense changes.

Mean intrahost AF/prevalence between NS and S mutations.

### 2.6.3 Intrahost variants distribution in coding genes

X out of 10 protein-coding genes of the viral genome.

Normalizing variants/ kb-gene-lenght (v/kbgk), the higher density was observed in ORF6 (16.21 v/kbgl).

Examples from literature: intrahost "15474¿T¿G" "28971¿A¿G" etc

### 2.7 Other metrics of polymorphic sites

-consensus vs non-consensus variants per position - how do samples with same consensus compare intrahost wise, is within this groups variance of AF smaller than in general?

## 3 Comparing intrahost SNVs with population-level SNVs from GISAID