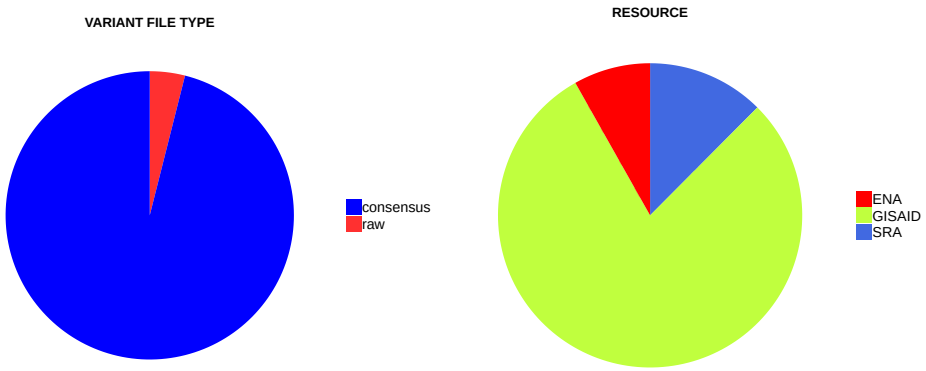
Viral Beacon Data freeze 20200524 - Variant files statistics

1 Summary stats

1. 38009 variant files

   by data type :
   intra-host=1497
   consensus=36512

   by data resource:
   SRA: 4650 (raw sequence files)
   ENA=3053 (consensus sequence files)
   GISAID=29629 (consensus sequence files)



   by resource & pipeline:
   Illumina SRA-galaxy=1497 (raw/intrahost variants files)
   ONT SRA-Medaka=3830 (consensus variants files)
   ENA=3053 (consensus variants files)
   GISAID=29629 (consensus variants files)

   by sequencing technology:
   SRA: Illumina=1497, Oxford Nanopore=3830,
   ENA, GISAID: (many, mixes, see by resource stats)

2. 37696(*) samples (SRA: 5260, ENA: 3053, GISAID: 29629) (*) This number is based on unique sample id and common samples that are cross-referenced across the databases.

   Note: Some variant files have the same samples as source: 24 SRA samples have more than one Illumina run, 17 SRA samples have more than one ONT run, and 26 SRA samples have both Illumina and ONT runs. There are 246 samples are common to SRA (ONT) and GISAID datasets. We don't know if some sample counts are also common to SRA (Illumina) and GISAID, or common to ENA and GISAID, or common to SRA and ENA.

2 Stats variant files metadata by resources (SRA, ENA, GISAID)

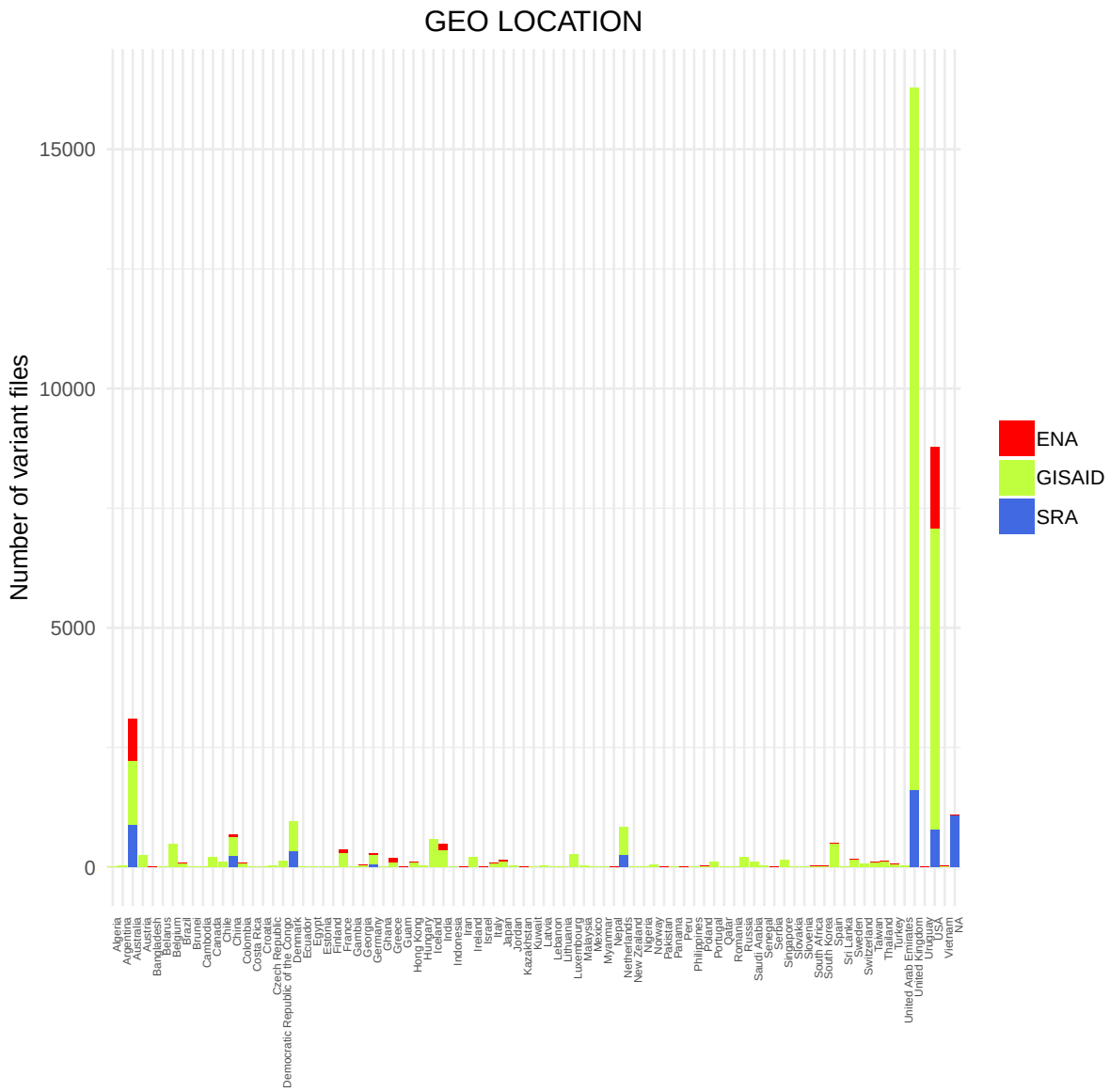


Figure 1: Distribution of variant files by country by resource

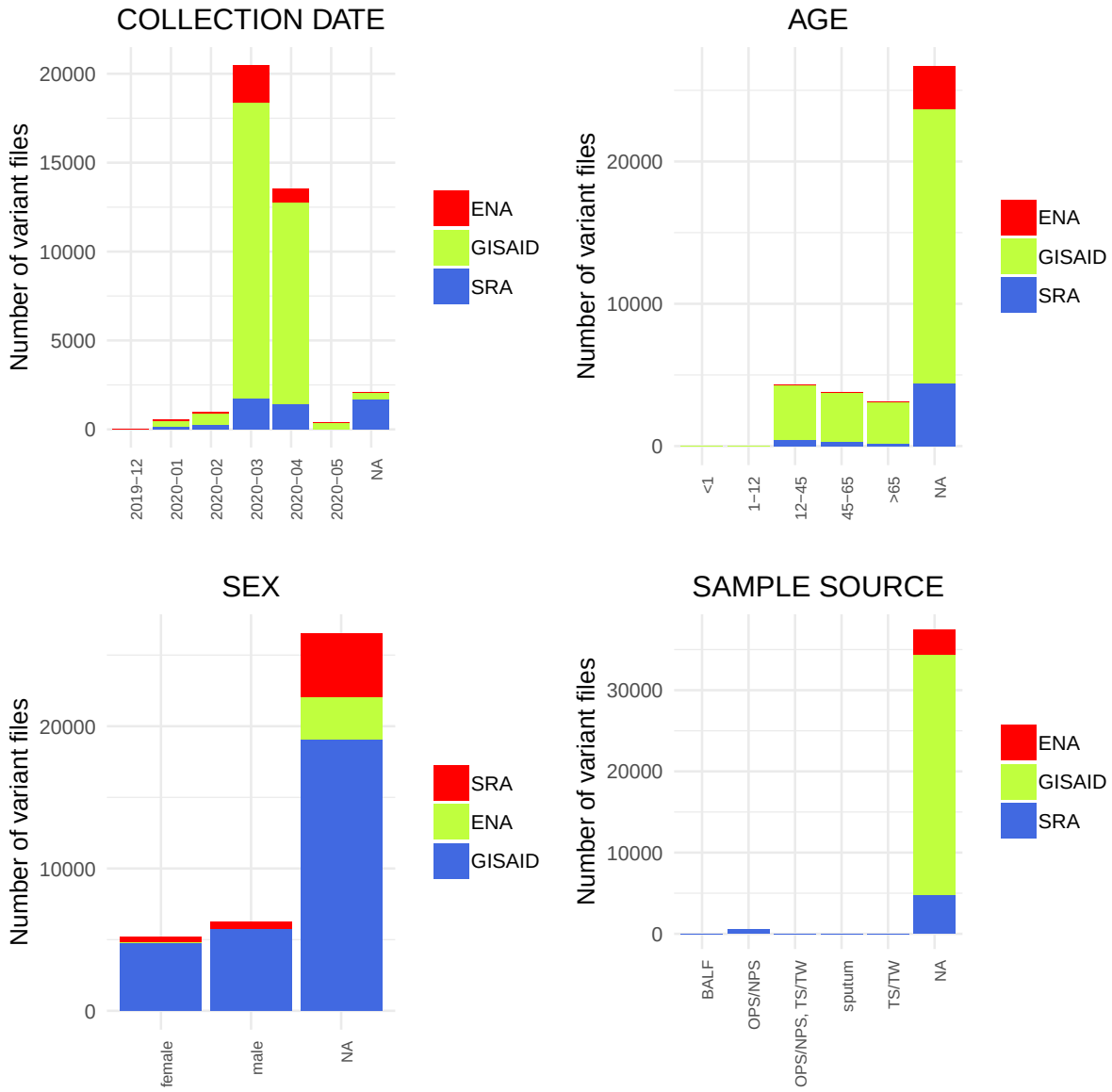


Figure 2: Distribution of variant files by collection date, host age, host sex, sample source by resource

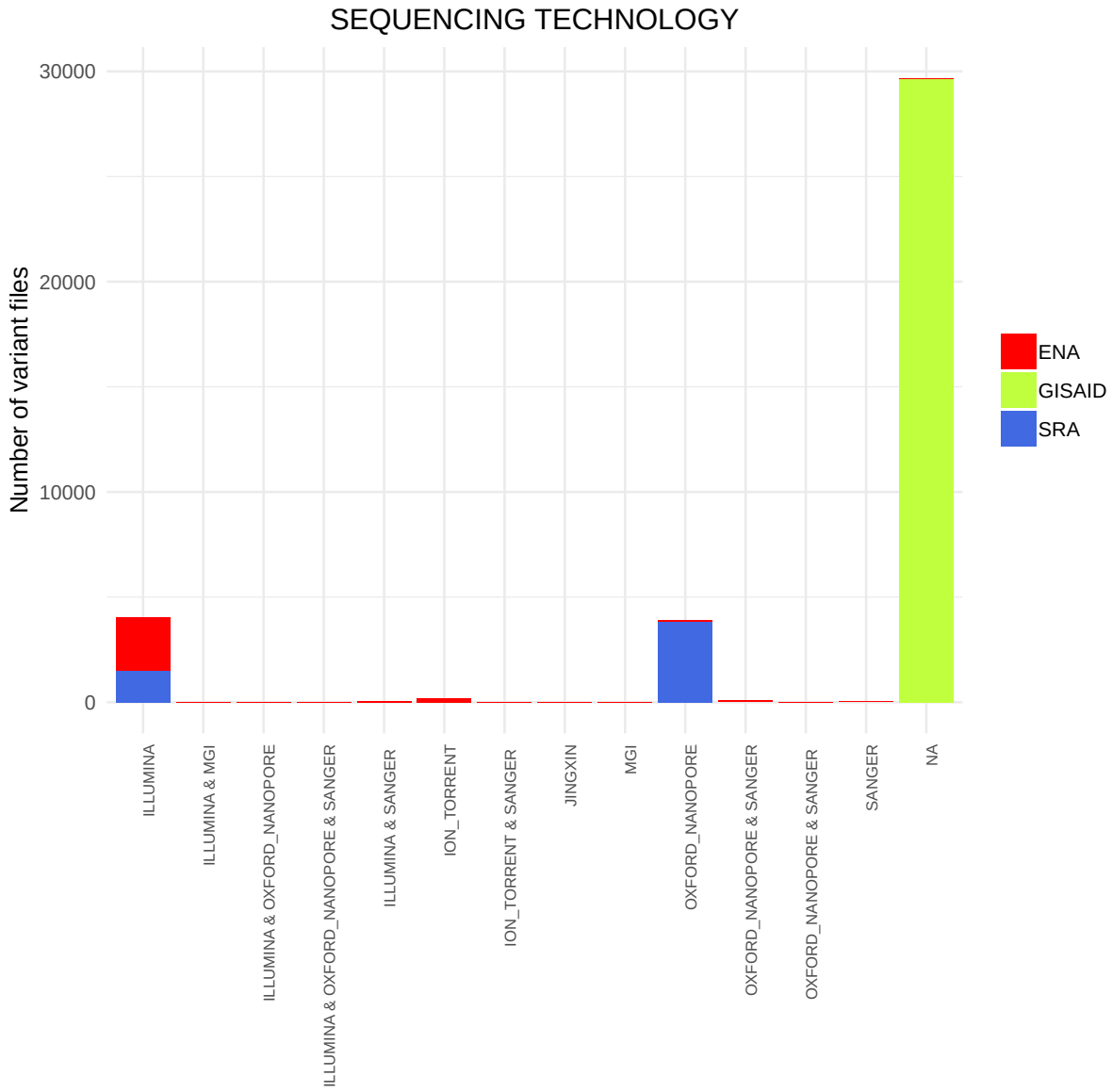


Figure 3: Distribution of variant files by sequencing technology by resource