locations_of_tags_for_primerid_data.pdf

Figure 3.1: The structure of Primer ID raw sequence data. A) cDNA structure with locations of PCR priming site, spacer sequences, MID and Primer ID. B) A forward strand Primer ID raw sequence read produced in PCR step. C) A reverse sequence produced in the PCR step. The primers tags the amplicon region which is being amplified, MID tags the specific individual from whom the sample is obtained and Primer ID tags the template RNA sequence. Two spacer sequences are used to separate the MID sequence from Primer ID and PCR priming site. The nucleotide sequence and length of spacers are customary and may vary between the projects.

Seq2Res_primerid_file_format.pdf

Figure 3.2:Different files required as input for the PIDA algorithm. Each column in all files has to be separated by a single tab. A) The amplicon specific primer file contains gene names that are amplified; the forward and reverse primers used for each amplicon; and the start and end nucleotide positions of the targeted amplicon region relative to standard HIV *pol* reference sequence. B) The MID file contains names of MID sequence used to tag the samples and sample names for identification. C) The gene file contains the amplicon names and minimum required sequence length for forward and reverse sequences. The information displayed in the files are just for the purpose and can be changed as required.

Note: x1, y1, x2 and y2 can be replaced with numbers that represent the amplicon start and end positions.

Seq2Res_primerid_pipeline_flow.pdf

Figure 3.3: The logical flow of the PIDA algorithm to process raw sequence data generated with Primer ID technology into generation of consensus sequence. The colors denote PIDA processing steps. Red: sequence demultiplexing, Green: Sequence filter by length, Blue: sequence filter by number of sequence represent by each Primer ID, Black: Quality trimming, Purple: Sequencing binning, Sky blue: Consensus sequence generation

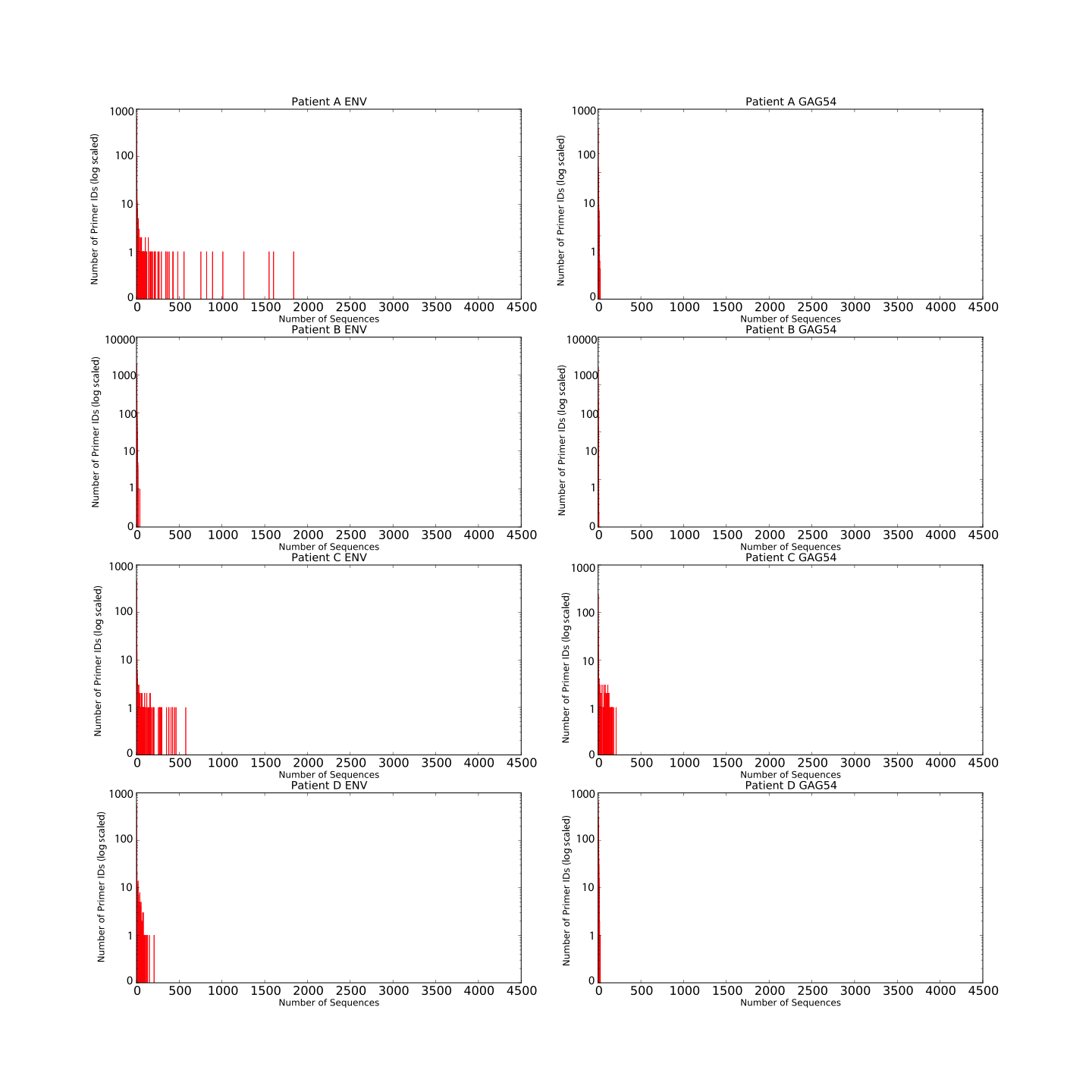


Figure 3.4: Count of Primer IDs representing specific number of sequence reads in *ENV* and GAG54 amplicons of Run1 Patients.

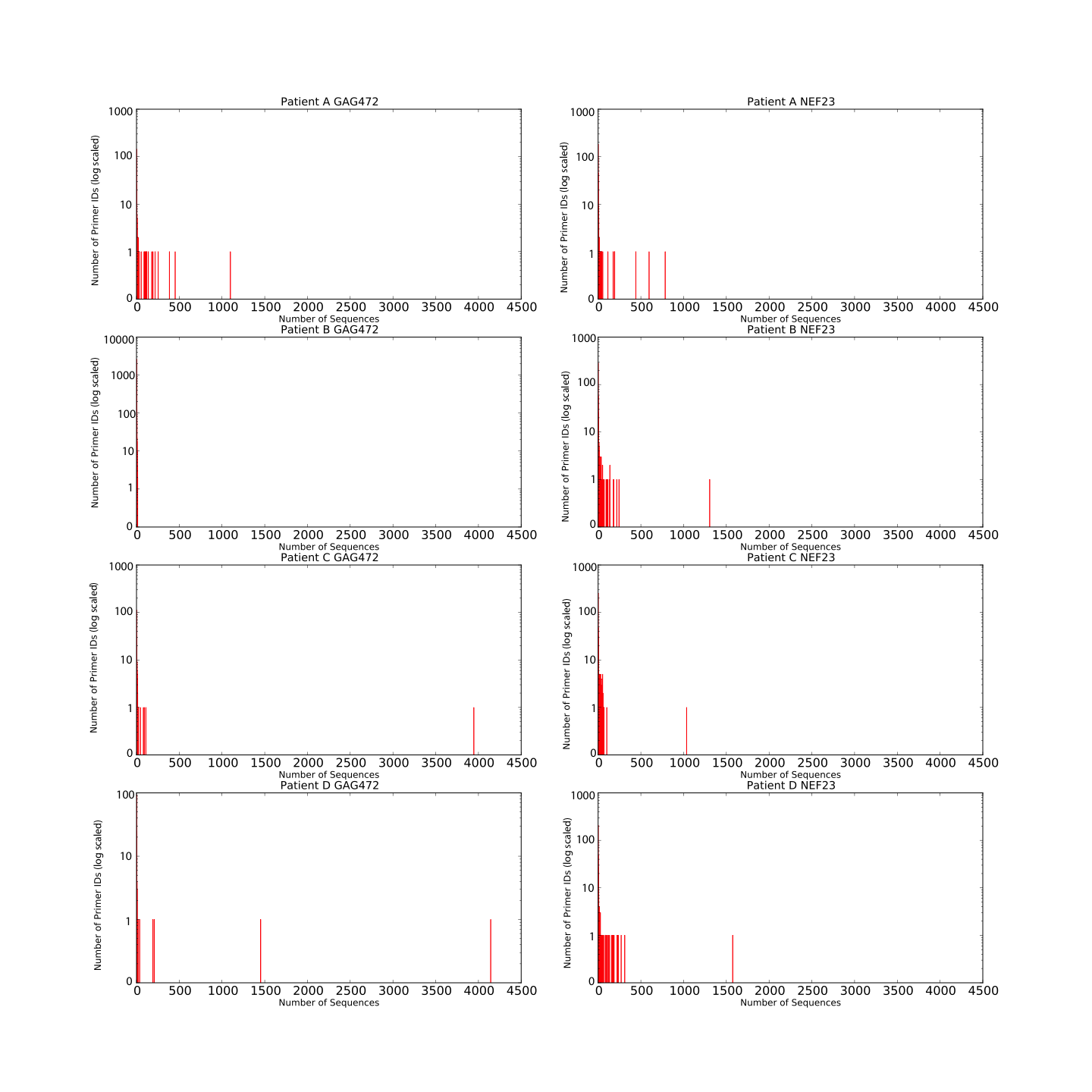


Figure 3.5: Count of Primer IDs representing specific number of sequence reads in GAG472 and NEF23 amplicons of Run1 Patients.

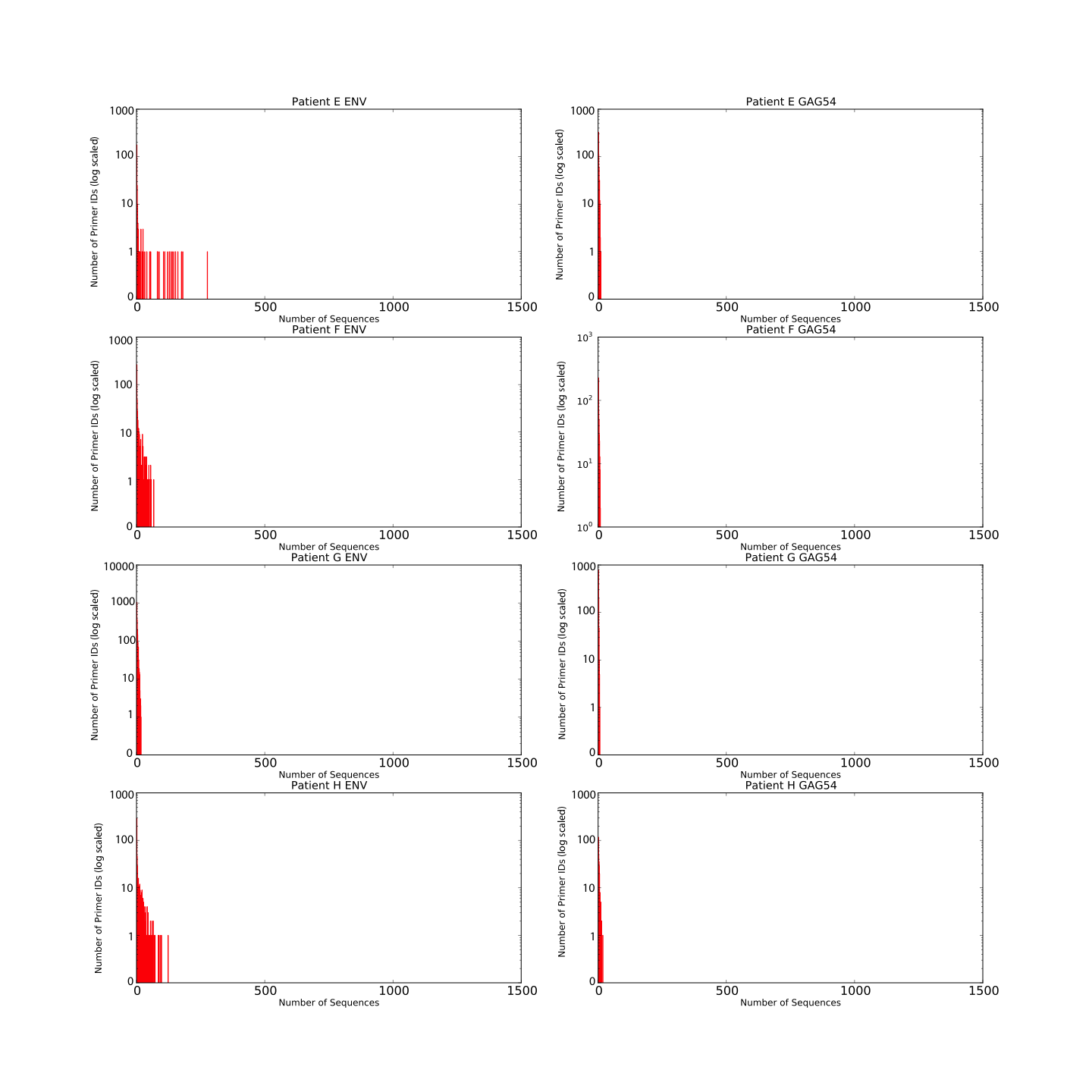


Figure 3.6: Count of Primer IDs representing specific number of sequence reads in *ENV* and GAG54 amplicons of Run2 Patients.

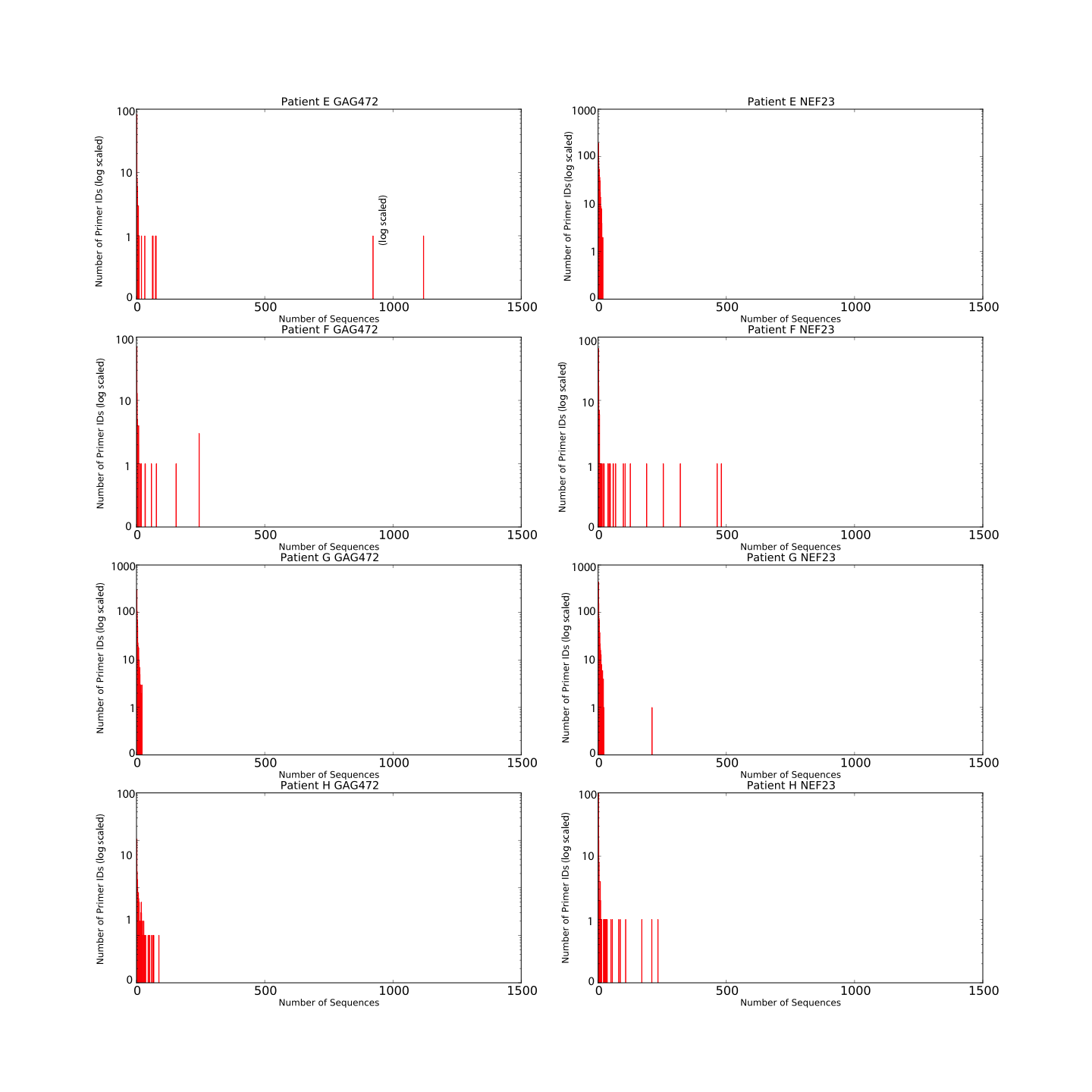


Figure 3.7: Count of Primer IDs representing specific number of sequence reads in GAG472 and NEF23 amplicons of Run1 Patients.

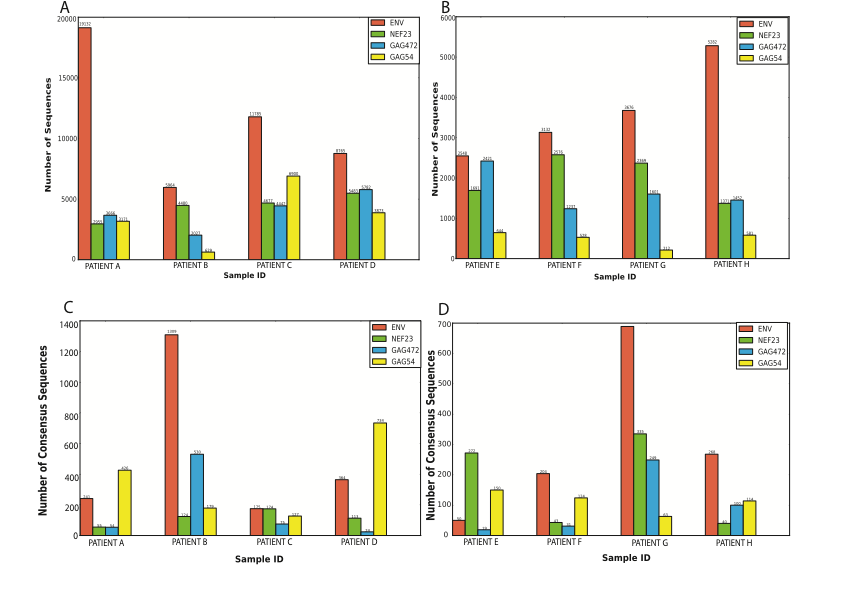


Figure 3.8: The distribution of final total number of sequences, each represented by Primer ID with greater or equal to threshold number of sequence, per amplicon per sample before consensus sequence generation in (A) Run1 and (B) Run2. The total number of consensus sequences generated per amplicon per sample in (C) Run1 and (D) Run2. Each consensus sequence is generated by collapsing the sequences represented by a unique Primer ID in an amplicon of a sample.