Virus discovery from high throughput sequencing data often follows a bottom-up approach where taxonomic annotation takes place prior to association to disease. Albeit effective in some cases, the approach fails to detect novel pathogens and remote variants not present in reference databases. We have developed a species independent pipeline that utilises sequence clustering for the identification of nucleotide sequences that co-occur across multiple sequencing data instances. We applied the workflow to 686 sequencing libraries from 252 cancer samples of different cancer and tissue types, 32 non-template controls, and 24 test samples. Recurrent sequences were statistically associated to biological, methodological or technical features with the aim to identify novel pathogens or plausible contaminants that may associate to a particular kit or method. We provide examples of identified inhabitants of the healthy tissue flora as well as experimental contaminants. Unmapped sequences that co-occur with high statistical significance potentially represent the unknown sequence space where novel pathogens can be identified.