Using pangenome variation graphs to improve mutation detection in a large DNA virus

Caroline Wright[1] (ORCID: 0000-0002-1770-4325), Chandana Tennakoon[1] (ORCID: 0000-0001-8614-6534), Lidia Lasecka-Dykes[1] (ORCID: 0000-0003-3188-4198), Tim Downing[1,*] (ORCID: 0000-0002-8385-6730)

[1]Pirbright Institute, Surrey, UK
*To whom correspondence should be addressed

Abstract

Accurately quantifying viral genetic diversity is essential for understanding pathogen evolution, transmission, and emergence. However, standard approaches that map sequencing reads to a single linear reference genome introduce substantial reference bias, particularly for samples that are divergent, recombinant, or belong to rare lineages. Pangenome variation graphs (PVGs) mitigate this issue by representing multiple genomes within a unified graph structure, enabling read mapping across all observed and potential haplotypes. Despite this, PVGs have rarely been applied to viruses. Here, we address this gap by constructing and evaluating the first PVG for lumpy skin disease virus (LSDV), an emerging poxvirus of global importance. We generated PVGs of different sizes and mapped Illumina datasets using Giraffe, benchmarking performance against linear reference mapping with Minimap2. A minimal three-sample PVG containing one representative from each major lineage recovered 97% of known LSDV nucleotide diversity while reducing PVG size by >95% relative to a 121-sample PVG. PVG-based mapping detected more SNPs than linear mapping, including variants at genes involved in host recognition and immune evasion, and identified lineage-specific mutations that improved subclade phylogenetic structure. Notably, 27% of SNPs discovered using PVGs could not be projected onto the linear reference, highlighting the extent of reference bias and the limitations of single-genome references. We propose a generalisable strategy for PVG construction that leverages viral population structure to maximise biological signal while minimising computational cost. Our findings demonstrate that PVGs substantially enhance SNP discovery in LSDV, with direct implications for genomic surveillance, outbreak tracing, and the detection of recombinant vaccine-related lineages in LSDV and other large DNA viruses.

Keywords: Virus, pangenome, pangenome variation graph, LSDV, pangenome graph, lumpy skin disease, poxvirus

Abbreviations

ADE, automated de novo error; AF, allele frequency; BAM, Binary Alignment Map; BQ, base quality; CDS, coding sequence; GAM, Graph Alignment Map; GBWT, graph Burrows–Wheeler transform; GFA, Graphical Fragment Assembly; GFF, general feature format; GTF, gene transfer format; GTR, general time reversible; IGV, Integrative Genomics Viewer; IQR, interquartile range; Kb, kilobase; LD, Lumpy skin disease virus gene designation; LSDV, Lumpy skin disease virus; MAFFT, multiple alignment using fast Fourier transform; MQ, mapping quality; NCBI, National Center for Biotechnology Information; ODGI, Optimized Dynamic Genome/Graph Implementation; PAV, presence–absence variant; PCA, principal component analysis; PGGB, Pangenome Graph Builder; PVG, pangenome variation graph; QC, quality control; RDAF, read-depth allele frequency; SAM, Sequence Alignment Map; SNP, single nucleotide polymorphism; Ti/Tv, transition–transversion ratio; VCF, variant call format; VG, Variation Graph toolkit; WGS, whole-genome sequencing; wfmash, weighted fast minimizer-based mash-like aligner.

Impact Statement

Genomic surveillance of viral pathogens typically relies on mapping reads to a single reference genome, a practice that systematically misses variation in divergent or recombinant isolates. We show that pangenome variation graphs (PVGs) can reduce this limitation in viruses. Using lumpy skin disease virus (LSDV) as a model system, we demonstrate that a compact PVG constructed from only three representative genomes captures most known genomic diversity, and substantially improves mutation detection compared with linear reference mapping. PVGs recover biologically meaningful variants at genes involved in immune evasion and host interaction, resolve closely related subclades, and reveal mutations that cannot be identified using a single reference genome. This work provides a generalisable population-structure-guided strategy for building reference viral PVGs that balance computational efficiency with sensitivity, offering clear benefits for genomic surveillance, outbreak reconstruction, and the detection of recombinant lineages.

Data Summary

All supporting code and protocols have been provided within the article or through supplementary data files. Supplementary tables are available at Figshare: https://figshare.com/s/7a879df5b5452d04585b with all code at https://github.com/downingtim/LSDV_pangenomics

Introduction

Traditional genomic analyses rely on mapping sequence reads to a single linear reference genome. However, this approach introduces reference bias when the sequencing reads differ from the reference, causing mismatched reads to align poorly and resulting in missed or mis-called SNPs, as well as structural and copy number variants [1-4]. Such biases reduce the accuracy of downstream phylogenetic reconstruction and genotype–phenotype association [5,6]. Importantly, neither consensus nor multi-sample linear references have overcome these limitations in viral systems [7,8].

Pangenome variation graphs (PVGs), also known as pangenome graphs or sequence graphs, address this reference bias by representing multiple genomes together in a single graph structure that encodes the nucleotide-level diversity of a sample collection [1]. PVGs have proven effective in human genomics, improving variant detection and reducing reference bias, particularly at structurally variable regions [1,9]. Their key advantage is that they reveal additional variants, including biologically meaningful changes that are invisible to linear reference approaches. Importantly, PVG-based read mapping is now computationally efficient, sensitive and accurate [1,9-13].

PVGs have seen limited use in viral genomics so far [14]. One notable exception is the PVG constructed for the circular 3.2 Kb hepatitis B virus, which successfully mapped more than 8% of reads that failed to align to linear references, particularly in highly diverse samples [10]. Viruses are relatively short, which reduces computational overheads, and recent PVG methods now support genomes of any length [15]. PVGs are well suited to viral datasets because they capture variation associated with phenotype, recombination, and rare divergent lineages, features that often pose major challenges for linear reference approaches. These advantages are especially valuable for tracking recombinant vaccine-derived lineages in livestock viruses.

Here, we demonstrate how PVG read mapping can improve inferences compared with mapping to a linear reference, using Lumpy skin disease virus (LSDV) as an exemplar. LSDV is one of three members of the Capripoxvirus genus within Poxviridae. It causes severe economic losses by infecting cattle, resulting in reduced milk production, hide damage, and mortality [16-18]. This large dsDNA virus is a notifiable disease by the World Organization for Animal Health (WOAH). Over the past two decades, LSDV has spread from sub-Saharan Africa into the Middle East, Asia, and more recently Southeast Asia and China [19-22]. LSDV is an emerging virus and accurate measurement of its genetic diversity is essential for linking evolution to geographic spread and drivers of socio-economic impact.

The LSDV genome is ~150 kb in length, AT-rich (75%), and typically encodes 156 genes [23]. It has three major recognised phylogenetic lineages: vaccine-related Clade 1.1, wild-type Clade 1.2, and recombinant Clade 2 [24-25]. Although LSDV evolves more slowly than RNA viruses, extensive recombination between wild-type and vaccine strains has generated mosaic haplotypes with altered phenotypes and possible vaccine-escape properties [26-28]. These mosaics cannot be fully represented by a single linear reference genome [29-30]. Furthermore, LSDV genomic diversity is elevated at genes involved in host recognition and immune evasion, which are typically located near the genome ends and require accurate variant detection. Such variation is critical for outbreak source tracing, monitoring recombinant lineages, and understanding mutations that influence cattle immune responses.

Here, we tested whether a PVG-based read mapping strategy improves variant detection compared with linear reference mapping, using LSDV as a model. We constructed PVGs from all available LSDV genome assemblies using existing tools [15] and mapped both real and simulated Illumina libraries to multiple PVG versions using Giraffe [12] and, for comparison, to a linear reference using Minimap2 [31]. We then used population structure to define a minimal yet representative PVG design that can be generalised to other viral systems. This approach is computationally efficient and enables rapid PVG-based analyses. We show that this reveals additional mutations and provides improved insights into LSDV genetic diversity from evolutionary, phylogenetic and functional perspectives.

Methods

Genome assembly collection, alignment, phylogenetic analysis, quality control and annotation

All available complete LSDV genomes (n=128) were downloaded from the NCBI nucleotide database [32] (15th November 2023), from which the first and last 300 bp of the genome were masked. This initial dataset was aligned with Mafft v7.453 [33] using automatic optimisation and default parameters (Figure 1).

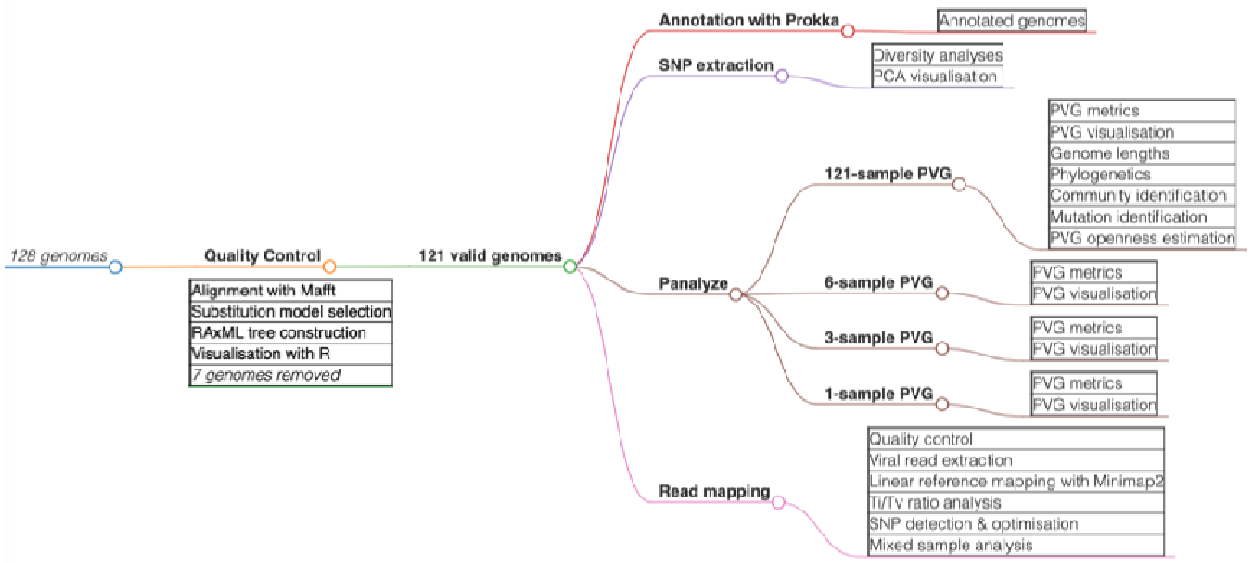


Figure 1. Methods summary showing the flow of data and analyses. After quality control of the initial dataset of 128 genomes, 121 validated genomes were subject to annotation, recombination detection, diversity analyses, PCA, phylogeny construction, PVG creation from 121-, 6-, 3- or 1-sample sample/s, and association PVG investigation (coordinate comparison, annotation, visualisation, community detection, and comparisons across genomic regions).

Evolutionary relationships among these genomes were reconstructed using RAxML-NG (Randomised Axelerated Maximum Likelihood) v1.2.0 [34] with a GTR (general time reversible) model and gamma substitution rate heterogeneity selected by modeltest-ng [35]. These phylogenies were midpoint-rooted and visualised using the following R v4.3.2 [36] packages: ape v5.7-1 [37], ggtree v3.8.2 [38], phangorn v2.11.1 [39], Rcpp v1.0.11 [40], RcppArmadillo v0.12.6.6.0 [41], phytools v2.0-3 [42] and treeio v1.24.3 [43].

Seven genomes were not included because their assemblies were reverse complemented [44], suggesting the veracity of the information may be low. Of these seven genomes, OK422493 (Ranchi 1 P30, 31/12/2019, India), OK422492 (Ranchi 1 P10, 31/12/2019, India) and ON400507 (208 PVNRTVU, 2020, India) were nearly genetically identical and were previously classified as LSDV Clade 1.2.2 (Breman et al 2023), and OM373209 (BH3 CHN 20, China) was previously classified as recombinant LSDV Clade 2.5.1 (Breman et al 2023). The three others were OM984486, OQ606832 (NIAB, PVNRTVU, India, 2022) and OQ427097 (West Bengal, 2022, India).

This resulted in a high-quality dataset of 121 genomes was annotated using Prokka v1.14.6 [45] for the Capripoxvirus genus using the isolate KX894508 (155920/2012) annotation that had 163 CDS regions (Figure 1). The median genome length was 150,575 bp; MW631933 (Morocco, 14/07/2017) was the

sole sample with a length <150 Kb. The relative difference in genome coordinates across the 121 samples was small: the median coordinate difference compared to KX894508 was 109 bp (Figure S1). MW631933 was the only sample with coordinates >360 bp different to the others.

Genome partitioning, SNP analysis, and phylogenetic resolution

Alignment and phylogenetic resolution of the 121 genomes were implemented as above (Figure 1). This was also applied for the core, 5' accessory and 3' accessory genomic regions. The core region was defined as a conserved 93 Kb region between bases 13,851 (the start of gene LD020, encoding a ribonucleotide reductase small subunit) and 106,910 (before gene LD116, encoding a RNA polymerase subunit),inclusive, as per previous work [24]. The 5' accessory genome extended to base 13,850 (13,610 bp in total). The 3' accessory genome was designated from base 106,911 onwards, with the core region at 13,611-106,910 bp.

The SNPs were extracted per sample as VCFs using SNP-sites v2.5.1 [46]. Comparisons and analyses of sequence diversity across groups and between sample pairs was implemented with VCF-kit v0.2.9 [47] and vcflib [48]. The core region had five triallelic sites, the 5' end had one such site and the 3' end had four triallelic sites. We performed principal component analysis (PCA) using R package adegenet v2.1.10 [49] based on genome-wide SNP data (Figure 1).

Pangenome variation graph creation

We used Panalyze [15] to implement PVG construction and analysis. We used it to make four PVGs: one with all 121 assembled genomes, another with three representing the three major clades, one with six (spread geographically and over time), and one with a single sample (KX894508) duplicated to differ at the first base only (Figure 1). KX894508 was used as the linear reference genome. Panalyze created PVGs in Graphical Fragment Assembly (GFA) format using the pangenome graph builder (PGGB) v0.5.4 pipeline [50] via all-to-all alignment with wfmash v0.7.0 [51]. Multiqc v1.14 [52] was used to collate and assess PVG metrics. These PVGs were converted with "convert" from variation graph (VG) v1.43.0 [9], indexed with VG's autoindex function and the corresponding genome files were indexed with SAMtools [53].

An advantage of using graphical formats is the ability to perform computationally scalable indexing through bidirectional graph Burrows Wheeler Transform (GBWT) indices, which store haplotype paths as nodes [13,54]. Consequently, we created GBWT-indexed versions of our PVGs using the corresponding haplotypes available with VG's gbwt function. We used the default minimiser length of 29 and window size of 11 for the Giraffe minimiser index. The GBWT indices were created using the greedy path-cover algorithm across all paths and were converted with VG. Then VG's snarls function was used to create a snarls file. This was used to create a distance index of the PVG's snarl decomposition so the minimum distances between nodes in the PVG were calculated. VG's minimizer function subsequently used these distances to identify each k-mer's position within the PVG for each haplotype, along with their corresponding PVG distances. VCFs were obtained from VG's deconstruct function using the GBWT-indexed graphs.

Pangenome variation graph manipulation, visualisation and annotation

Panalyze analysed these PVGs using ODGI (optimized dynamic genome/graph implementation) v0.8.3 [55], including quantifying numbers of SNPs, indels and compound mutations using VG's "deconstruct" function. Panalyze used gfautil v0.3.2 [56] to get mutation rates and coordinates for genomes at each path and site. Subgraphs were selected using gfatools v0.4 [57], evaluated with gfastats v1.3.6 [58], analysed with gfastar v0.1 [58], and visualised with waragraph [59] as well as VG. Presence-absence variants (PAVs) were identified using ODGI via Panalyze. Panalyze visualised the PVGs with ODGI [55] and the R package ggplot2 v3.4.4 [60].

The PVGs were annotated using Prokka [45] as above and were indexed with SAMtools v1.19 [53]. Using the reference genome annotation for KX894508, Busco v5.5.0 [61] found a mean of 155.8 complete unique CDSs per genome (standard deviation 1.5). The Prokka GFFs were converted to gene transfer format (GTF) with gffread [62] and parsed. The annotation position information was merged with the ODGI [55] rendering of the PVGs, creating one CSV annotation file per PVG, which was visualised using Bandage v0.8.1 [63].

Community detection and visualisation

PVG communities reflected the relative similarity of genomes based on pangenome data for a specific region. Here, these were identified via Panalyze's implementation of pairwise alignment using wfmash [51] via PGGB [50] with a k-mer of 19 bp and a window size of 67 bp, subject to thresholds of 80% sequence mash-level identity and ten mappings per node. Panalyze indexed the file names using fastix [64], and evaluated these communities with PGGB scripts [50]. This built a network with the contigs as nodes and mappings as weighted edges, such that the weights were proportional to the percentage identity of the mapping and its length. These networks were visualised with igraph tools [65].

Read library sample collection and quality control

76 Illumina read libraries were downloaded from the SRA (24th May 2024) and six additional unpublished ones were added to make a dataset of 82 libraries. The metadata for each library was retrieved where available, including host species, year/country of isolation, sequencing platform, library source/selection and preparation approach (Table S1). We also analysed ten simulated paired-end read libraries based on the genomes of all samples using art_illumina v3.19.15 [66]. These ten libraries had varied error rates scaled from 100% of the normal rate to 90%, 80%, 70%, 60%, 50%, 40%, 30%, 20% and 10% of the normal rate.

Adapter trimming, removal of reads with low base quality (BQ) scores (phred score <30), exclusion of ambiguous (N) bases, correction of mismatched bases in overlapping regions, and trimming of poly-G tracts at 3′ ends was completed using Fastp v0.23.4 [67]. Low-quality bases with BQ <30 and reads with lengths <110 bp were removed using the FASTX-Toolkit v0.0.13 [68]. FastQC v0.12.1 [69] and MultiQC v1.0 [52] were used to assess the effectiveness of Fastp and the FASTX-Toolkit. 89% of Illumina read library bases had BQ>30, similar to previous work [70]. The number of screened reads was higher for metagenomic libraries (mean 36,146k) compared to WGS (3,063k) and amplicon (678k) libraries (Table S2). The fraction of reads excluded because they were low quality or had too many Ns was similar across library types. Reads that were too short or adapter-trimmed were only in WGS libraries. The median read duplication rates were higher for amplicon (19.0%) and WGS (18.8%) libraries compared to metagenomic ones (10.9%).

We estimated species-level abundances for each read in each library using Kraken2 [71] with the core nucleotide database. All libraries had reads stemming from *Bos*, *Cervus*, *Homo* and *Ovis* in addition to LSDV (online Supplementary Material 1). The relative numbers of reads allocated to each stage were assessed (Table S3).

Pangenome variation graph read mapping, testing and comparisons

We mapped the viral reads of each library using Giraffe in VG v1.43.0 [9] to each PVG to generate Graph Alignment Map (GAM) files. Giraffe mapped each read library to the PVG containing one (named Giraffe_1 here), three (Giraffe_3), six (Giraffe_6) and all- (Giraffe_all) samples. VG uses efficient PVG indexing and read mapping using a partial order alignment [2]. The PVGs were initially in XG format (topological representation) with accompanying GCSA2 indices for read mapping using VG-MAP. To enable mapping with Giraffe, the PVGs were converted from GFA to GBZ format,

which includes a compressed graph and an index of all haplotypes. The GBZ file was indexed with VG, and the minimizer distances were used by VG's call function. VG's surject function was used to create Binary Alignment Map (BAM) files, which represent the reads at their reference nodes in the PVG. We obtained GAM read mapping statistics across all snarls using VG's stats function. The VG paths and VG mod functions were used to extract haplotype paths from the individual graphs. Mapping with Giraffe was much better than VG-MAP [13], so the latter is not discussed further.

We used GAM files to visualise read mapping patterns of individual samples on the PVGs. Augmented graphs were then generated using VG's augment function, with the following thresholds: a minimum path coverage of three, a base quality (BQ) of at least 10, and a minimum mapping quality (MQ) of 10. Snarls were interpreted based on the augmented graphs using VG's snarls function. VG's pack function generated PACK files from the augmented graphs, which are compressed coverage indices, analogous to Sequence Alignment Map (SAM) files. Each unique snarl in the PVG was used to screen for potential mutations in VCF files by VG call function, which used the PACK file and augmented graph data. Mutations were detected if the sample's paths in the PVG traversed the allele of interest. These PACK files were then used as input for VG's depth function to calculate path depth at regions with a coverage of six-fold or more. The compute resources and time taken for this VG-based approach was notably longer than that using the surjected BAM files. To identify candidate mutations from GAM files via VG, we assumed haploidy and required variants to have a MQ of 10+. This was processed by using the VG augment, snarls, pack and call functions to generate VCF files for valid candidate mutations (labelled with "PASS"). These VCF files were used to interpret SNPs with allele frequencies between 5% and 90%. The mapped reads in the GAM files were visualised with SequenceTubeMap [72].

We used Minimap2 v2.24-r1122 [31] to map libraries to the KX894508 reference genome and generate SAM files, which were then converted to BAM format using SAMtools v1.11 [53]. The latter was then used to sort these BAM files, including surjected BAMs, extract information on the mate coordinates, calculate insert size, resolve missing read pairs, remove duplicate reads, index them and extract read mapping statistics. Mutations detected with VG were extracted from the corresponding PACK files. BAM file metrics were obtained using SAMtools' [53] flagstat function, while read depth data were derived using SAMtools' [53] coverage function and was parsed with R to measure read depth, BQ and MQ.

Mutation detection and screening

We filtered for candidate mutations (SNPs and indels) based on the BAM files using BCFtools v1.20 [53] and Freebayes v1.3.7 [73]. BCFtools detects mutations based on the frequency of derived alleles in sequencing reads and has been shown to outperform alternative options like GATK in certain contexts [74]. Freebayes identifies mutations across haplotypes using a Bayesian approach [73]. Previous studies have shown that: stringent thresholds are effective for viral sequence data [75], both BCFtools and Freebayes are effective for mutation detection [76], and these tools work well with PVG-based approaches [77].

We focused on high-quality mutations that met the following criteria: BQ (QUAL) score of 30+, a minimum read depth of 10, at least five reads supporting the alternate allele, and whose location was not within 2.5 Kb of the genome ends. These thresholds were consistent with previous studies. For variant calling with BCFtools, we also required that variants were haploid. For Freebayes, we also required an allele frequency >0.05. These parameters were selected and optimised by manually inspecting candidate SNPs using IGV [78]. All resulting VCF files were normalised with BCFtools to represent multiallelic sites correctly. This process produced a VCF file for each sample across the mapping approaches (Giraffe and Minimap2), PVG options (one-, three-, six-, all-sample) and variant callers (BCFtools and Freebayes), generating 10 VCF files per sample. The correlation between SNP

read depths and their read-depth allele frequencies (RDAFs) was low (r2=0.03), indicating that read depth was not a limiting factor in our dataset.

Transition-transversion (Ti/Tv) ratios were calculated using BCFtools for all detected mutations before and after filtering. A neutral Ti/Tv ratio=0.5 was expected for non-conserved regions because there are twice as many mutational options for transversions compared to transitions. A Ti/Tv ratio > 2 can be indicative of functionally active or constrained regions, as observed in *Drosophila* [79] and humans [80-82]. A Ti/Tv ratio > 3 has been observed in highly constrained loci [83] including inter-species comparisons of viral genomes [10]. Although reduced Ti/Tv ratios can result from mutation saturation, here these effects may be less pronounced because we examined a slow-mutating dsDNA virus. To help identify reliable SNP-calling pipelines, we required that any valid mapper-PVG-caller combination have a Ti/Tv > 2.

Comparisons between SNP datasets were implemented using the BCFtools' concat function. We used the FLO toolkit (https://github.com/wurmlab/flo [84], SAMtools [53] and the Picard "LiftoverVcf" function [85] to transfer coordinates between VCF files with different reference genomes. This liftover converted variant calls from the three-, six- and all-sample PVGs to the KX894508 annotation, requiring a sequence identity of >98%. We identified valid mutations with mismatched reference alleles, which indicated that the KX894508 reference lacked the alternative allele. These "unlifted" mutations represented variants uniquely detected using PVG-based methods that were missed by traditional linear reference-based approaches. SNP data was processed with Python and visualised using R package ggplot2 v3.4.4 [60].

The functional impact of SNPs was examined by comparing the FASTA and GenBank annotations of KX894508, OQ511520 and KX764645 against the VCF files using an LSDV-specific variant effect predictor. This detected SNPs as causing nonsynonymous, synonymous, intergenic, stop-loss or stop-gain mutations within Clades 1.1, 1.2 and 2. We also calculated the rates of nonsynonymous and synonymous variants per sample.

Co-phylogenetic analysis of different mapping approaches' effects

We examined the phylogenetic and co-phylogenetic effects of SNP detection across different mapping strategies, directly comparing linear reference and PVG-based approaches. This used the merged SNPs from Giraffe_1 and Giraffe_6 (called Giraffe_1&3), and separately the merged SNPs from Giraffe_1 and Giraffe_6 (called Giraffe_1&6). To do this, we created a concatenated SNP alignment containing the valid SNPs for 78 samples across the six genomes in the six-sample PVG (total lengths 903.7 Kb). Four divergent samples (SRR8442780, SRR19090746, SRR19090747, SRR19090748) were excluded. Maximum likelihood phylogenies were inferred using RAxML-NG v1.2.0 [34] using a GTR+G4 model. The trees were mid-pointed rooted and visualised in R. We used phytools v2.0-3 [42] for cophylogenetic comparisons. The clades were annotated based on published data [24]. These alignments contained 1,181, 1,041 and 1,003 parsimony-informative sites for mapping to the linear reference with Minimap2, mapping with Giraffe to the one- or three-sample PVGs, and mapping with Giraffe to the one- or six-sample PVGs, respectively.

Results

Pangenome variation graph analysis of LSDV genomes

We constructed a PVG of 121 LSDV genomes (called the all-sample PVG) with Panalyze [15]. This PVG revealed varied genetic diversity across the genome with a two-fold higher SNP density at the 5' and 3' terminal ends compared to the core region (Figure 2). This was supported by a higher Ti/Tv ratio at the core (Table 1). Additionally, we found that the allele frequency spectrum decayed exponentially with sample count (Figure S2), indicating that most SNPs were rare and lineage-specific, with only a minority reaching frequencies > 50% across the three major clades.

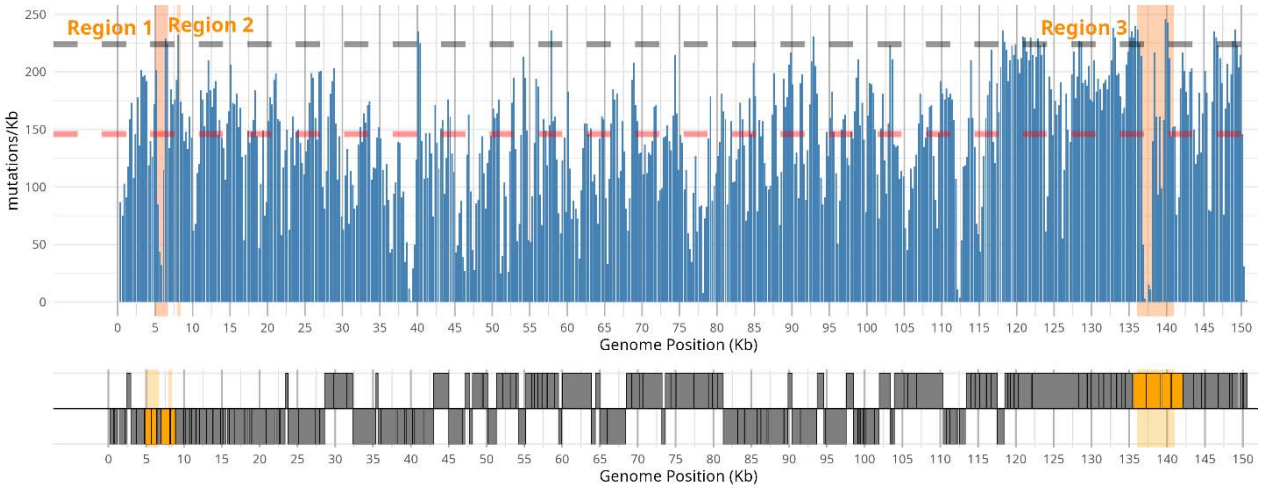


Figure 2. PVG-wide diversity (SNPs/Kb) in LSDV using 500 bp sliding windows (blue bars). The KX894508 genome annotation is represented below as grey boxes that are read in the forward (lower strip) or reverse (upper strip) directions. The red dashed line indicates the median number of SNPs/Kb. The black dashed line denotes the 95% percentile. Three regions of interest and associated CDSs are in orange: (Region1) 5,200-6,500 bp, (Region2) 8,055-8,231 bp and (Region3) 136,200-140,400 bp.

Region	Length (bp)	SNPs	Theta/Kb	Pi/Kb	SNPs/Kb	Ti/Tv
5' accessory	13,610	325	4.1	6.8	23.9	2.40
core region	93,059	1,076	2.3	3.8	11.6	3.49
3' accessory	43,644	1,107	4.7	7.6	25.4	2.37
total	150,313	2,436	3.3	5.4	16.2	2.89

Table 1. Genetic diversity across the whole genome, 5' end, core region, and 3' end. Values include the transition/transversion ratio (Ti/Tv), Watterson's theta per Kb (Theta/Kb), and the mean number of pairwise differences per Kb (Pi/Kb) across samples.

Subset-derived pangenome graphs capture broader LSDV diversity

Phylogenetic reconstruction of the PVG's SNPs recapitulated the known population structure [27] (online Supplementary Material 2), with three major clades: vaccine-related Clade 1.1, wild-type Clade 1.2 and recombinant Clade 2, along with several rarer lineages (Figure S3). We hypothesised that LSDV's simple population structure would allow a PVG based on a subset of samples. To explore this, we created PVGs from three and six genetically representative samples (referred to as the three-

and six-sample PVGs) (Table 2). The three-sample PVG had a representative from each of the main clades: KX764645 from vaccine-related Clade 1.1, KX894508 from wild-type Clade 1.2, and OQ511520 from recombinant Clade 2. Pairwise comparisons showed 2,943 SNPs between KX764645 and KX894508, 2,121 between KX764645 and OQ511520, and 1,973 between OQ511520 and KX894508. These values were consistent with the wider dataset, for which the median pairwise SNP distance was 1,858 (standard deviation 1,080). For the six-sample PVG, more than one sample per clade was used to account for assembly artefacts, differences in genome completeness, and subclade-specific diversity. Within Clade 2, MW732649 and OQ511520 differed by 1,272 SNPs, and MW732649 had novel variants absent from the other samples. In Clade 1.2, OK318001 differed from KX894508 by 706 SNPs and from KX683219 by 907 SNPs, and KX894508 and KX683219 differed by 611 SNPs.

Accession	Sample name	Subclade	Country	Year	PVGs
KX894508	155920/2012	1.2.1	Israel	2012	3 & 6
OQ511520	Thailand/Trang/2022	2	Thailand	2022	3 & 6
KX764645	Neethling-LSD_vaccine-OBP	1.1	South Africa	NA	3 & 6
OK318001	V281	1.2.3	Nigeria	2018	6
KX683219	KSGP_0240	1.2.2	Kenya	1974	6
MW732649	HongKong/2020	2	Hong Kong	2020	6

Table 2. GenBank accession and associated sample details for genomes included in the three- and six-sample PVGs.

The all-sample PVG had a total length of 155,008 bp (151,506 bp excluding non-ACGT bases), of which 147,614 bp were shared, indicating that ~97% of the LSDV genome was conserved. By comparison, the six-sample PVG was 151,246 bp (150,643 bp without Ns), indicating that only 863 bp (0.6%) of additional variation came from the extra 115 genomes. Graph complexity was substantially lower in the six-sample PVG: the sum of sequence lengths was 18,213 Kb in the all-sample PVG versus 904 Kb in the six-sample PVG. Moreover, the six-sample graph retained 82% of the all-sample PVG's unique nodes and 81% of its edges (Table 3). Similarly, the three-sample PVG (151,760 bp; 150,557 bp without Ns) was only 86 bp shorter than the six-sample PVG, retaining 94% of its nodes and 93% of its edges. The all- and six-sample PVGs contained 24% and 6% more SNPs than the three-sample PVG, respectively. Overall, these results illustrated that the three- and six-sample PVGs captured most LSDV genomic diversity and required substantially less computational time for analyses than the all-sample PVG.

Performance of PVG versus linear reference mapping across library types

We examined all available Illumina read libraries following careful quality control and viral read extraction to compare the outputs of mapping to these PVGs versus a single reference genome (Table S3, online Supplementary Material 3). Reads were mapped to a linear reference (KX894508) with Minimap2, and to the (pseudo-)one-, three-, six and all-sample PVGs using Giraffe. The one-sample PVG was composed of KX894508 duplicated. We examined the numbers of reads mapping in the resulting GAM and BAM files in the amplicon (n=27, Figure S4) and WGS libraries (n=46, Figure S5) and found that read libraries could be mapped well with either PVG-based or linear mapping approaches (Table S4). The differences between the median Minimap2 mapping rate and the four PVG options was negligible for amplicon libraries (all IQRs 99.5-99.9%). For WGS samples, Minimap2 had a slightly higher median (73.6%) compared to the PVG options (medians 69.0%-73.0%) in this more heterogeneous dataset. The simulated read libraries mapped better with Giraffe to the three- or six- or all-sample PVGs than to a linear reference with Minimap2 (99.9% vs 98.8%) (Figure S6). Metagenomic library mapping rates were much lower overall: they had an IQR of just 17-35% based

on linear read mapping and 16-32% from PVG mapping (Figure S7). Five of the six metagenomic libraries mapped better using linear reference mapping with Minimap2, with considerable heterogeneity across samples.

Property	PVG type		
	All	Six	Three
Length	151,506	150,643	150,557
Mutations	2,330	1,996	1,881
Nodes	7,207	5,903	5,520
Edges	9,836	7,939	7,397
Paths	121	6	3
Summed genome lengths (Kb)	18,213.4	903.7	451.9

Table 3. Properties of the all-sample ('All'), six-sample ('Six'), and three-sample ('Three') PVG. The length excludes bases denoted as "N".

Enhanced SNP detection using PVG-based Giraffe mapping

Giraffe produced equivalent or superior SNP detection to Minimap2 when mapping to the one-sample PVG (online Supplementary Material 4, Table S5). Giraffe identified more non-singleton SNP sites missed by Minimap2 than vice versa (28 vs 18), yielding 217 additional SNPs across the samples versus 125 SNPs unique to Minimap2. Moreover, Giraffe did not miss any SNPs detected by Minimap2 and found more SNPs per sample (median 5, IQR 4-6) than Minimap2 (median 3, IQR 3-4) in the 21 amplicon libraries with <100 SNPs (Figure 3). Most SNPs in the amplicon samples with >1,000 SNPs were shared by both methods: unique SNPs constituted only 0.2-0.4% per sample.

Subconsensus SNP detection was higher for the PVG-based mapping approach (Figure S8, Figure S9, online Supplementary Material 5). Moreover, a median of 27% of SNPs from the three-sample PVG could not be transferred to the linear reference KX894508, reflecting allele differences between reference sequence and observed states (Table S6). SNPs were evenly spread among the genomes from the three-sample PVG (Figure 4A), but more skewed in the genomes from from the six-sample PVG, suggesting that the three additional genomes (OK318001, KX683219, MW732649) added limited extra information in the latter (Figure S10).

To accurately compare PVG-based mapping to linear mapping, we merged the detected SNPs from Giraffe mapping to both the one- and three-sample PVG, and the same for the one- and six-sample PVG. The merged one- and three-sample SNP data captured far more non-singleton SNP sites missed by linear mapping than vice versa (353 vs 24 sites, Figure 4B), amounting to 2,658 additional mutations across samples compared to 176 mutations uniquely associated with linear mapping (Table S7). Similar trends were also seen with the merged one- and six-sample SNP data dataset (Table S8). On average, more SNPs were unique to the merged PVG outputs (77-80 per sample) than to Minimap2 (~3 per sample) where ~267 SNPs on average were shared between the two approaches. In WGS libraries, SNP detection was higher in the PVG-based mapping (2,279 and 2,090 in the merged three- and six-sample PVG outputs, respectively) compared to Minimap2 (1,917). The same pattern was observed in amplicon, simulated, and metagenomic libraries.

We examined the level of non-singleton SNPs at three regions of elevated variation at 5,200-6,500 bp, 8,055-8,231 bp and 136,200-140,400 bp (Figure 5). Region 1 spanned LD008 (encoding a soluble interferon gamma receptor) and LD009 (encoding an alpha amanitin-sensitive protein) and had seven SNPs at 5,831-5,897 bp (0.105 SNPs/Kb) (Figure S11). Region 2 was at LD011 (encoding a CC chemokine receptor-like protein) that had three SNPs at 8,062-8,131 bp (0.042 SNPs/Kb) (Figure

S12). Region 3 spanned LD144 (encoding a kelch-like protein), LD145 (encoding an ankyrin repeat protein) and LD146 (encoding a phospholipase D-like protein) and had five SNPs at 136,626-712 bp (0.057 SNPs/Kb) and three at 140,272-297 bp (0.11 SNPs/Kb) (Figure S13). In each case, Giraffe mapping to PVGs identified more non-singleton SNPs missed by Minimap2 than vice versa: 14 vs 2 at Region 1, 6 vs 0 at Region 2, and 3 vs 2 at Region 3 (Figure 4C).

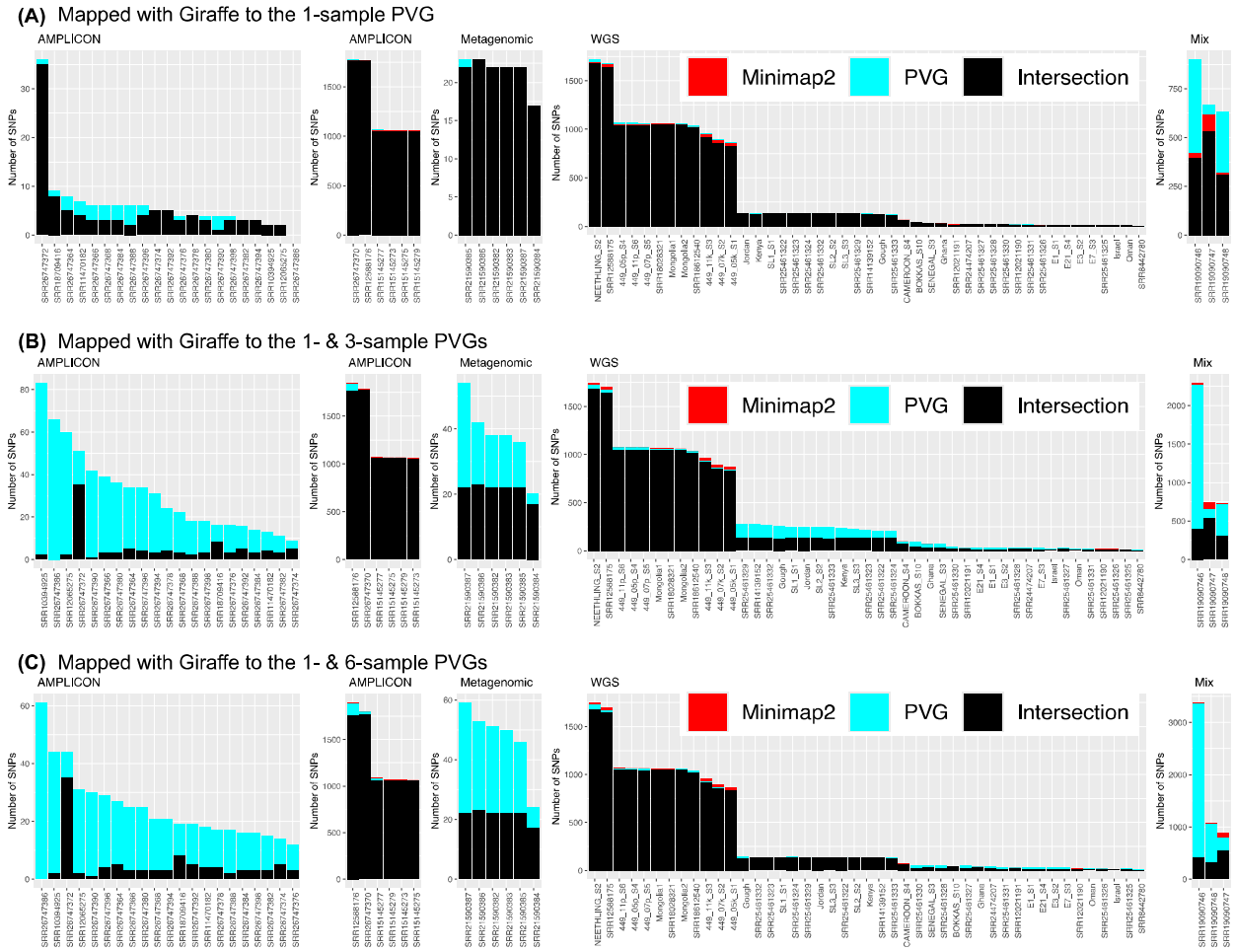


Figure 3. Number of SNPs detected by linear reference mapping with Minimap2 and Giraffe mapping to (A) one-sample PVG, (B) combined one- and three-sample PVGs and (C) combined one- and six-sample PVGs. Red colouration indicates SNPs captured by Minimap2 alone. Cyan indicates those captured by PVG-base mapping only. Black indicates those captured by both Minimap2 and PVG mapping. Data is arrayed across library type: simulated, amplicon, metagenomic and WGS. The y-axes are different for each dataset, and those for amplicon datasets are split to match their corresponding SNP ranges. Three mixed samples ('Mix') are shown separately from the other WGS samples.

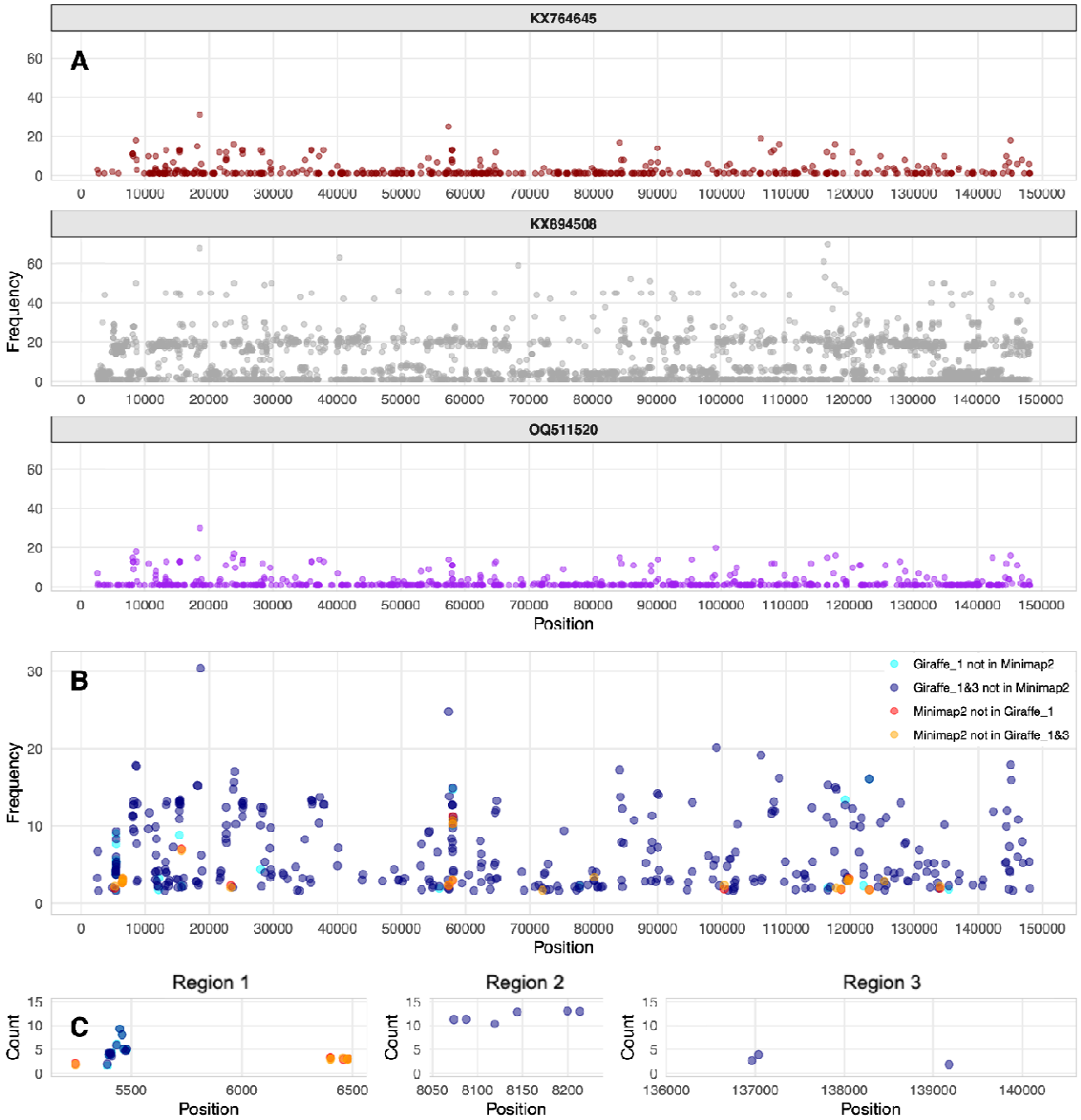


Figure 4. (A) Genome positions of mutations in the merged output of Giraffe mapping to the one- and three-sample PVGs showing the KX764645 (wine, top), KX894508 (grey, middle) and OQ511520 (purple, bottom) genomes. The bin size was 250 bp. (B) The number of novel non-singleton SNPs detected by mapping approach across the genome. Points indicating unique non-singleton SNPs missed by linear mapping but found by Giraffe mapping to the one-sample PVG (28 SNPs, cyan) and the merged one- and three-sample PVG outputs (352 SNPs, blue) at Regions 1, 2 and 3. Points show unique mutations found by linear mapped but missed by Giraffe mapping to the one-sample PVG (18 SNPs in red) and the merged mapping to the one- and three-sample PVGs (23 SNPs in orange).

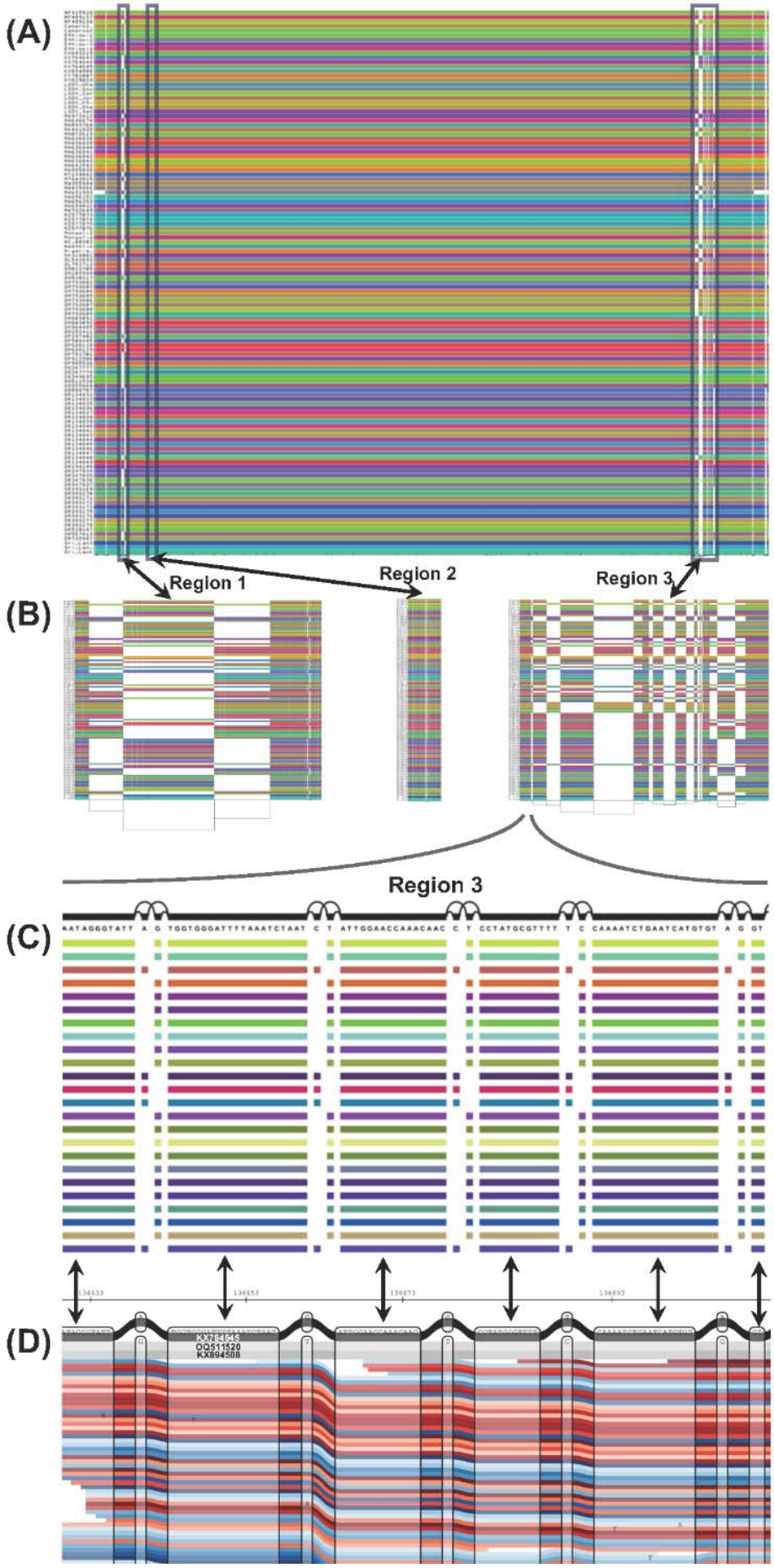


Figure 5. Visualisation of the LSDV PVG and the three identified regions of elevated variation where each panel shows increasing detail. (A) ODGI image of the complete PVG. (B) ODGI images of Region1-3, left: 'Region1' (5, 200-6,500 bp), middle; 'Region2' (8,055-8,231 bp), and right: 'Region3' (136,200-140,400 bp). (C) VG visualisation focussing on a region of LD144 (136,626-136,712 bp) within 'Region3', with five SNPs indicated. The arrows between (C) and (D) denote the matching regions between these panels. (D) SequenceTubeMap image of reads from LSDV sample 449_07p_S5 in red and blue mapped to the three-sample PVG across the same region of LD144 (136,626- 136,712 bp), showing the same five SNPs, where KX894508 (Clade 1.2) is in light grey, OQ511520 (Clade 2) is in grey, and KX764645 (Clade 1.1) is coloured black. The nodes contain sequences, and the edges join these nodes across the reads.

Phylogenetic and functional effects of novel SNPs from pangenome variation graph mapping

Phylogenies constructed from 78 read libraries using Minimap2 mapping to a linear reference or Giraffe to the combined one- and three-sample PVGs or one- and six-sample PVGs were similar (Figure 6). All identified several unique lineages with long external branches that were more basal, symptomatic of recombination and phylogenetic uncertainty. Giraffe mapping to the combined one- and three-sample PVGs displayed more population structure compared to mapping to the combined one- and six-sample PVGs or to a linear reference, such as subclade 1.2.2. SNPs ascertained with Minimap2 or Giraffe mapping combined one- and six-sample PVGs showed no clear differentiation of groups within Clade 1.2, unlike Giraffe mapping combined one- and three-sample PVGs.

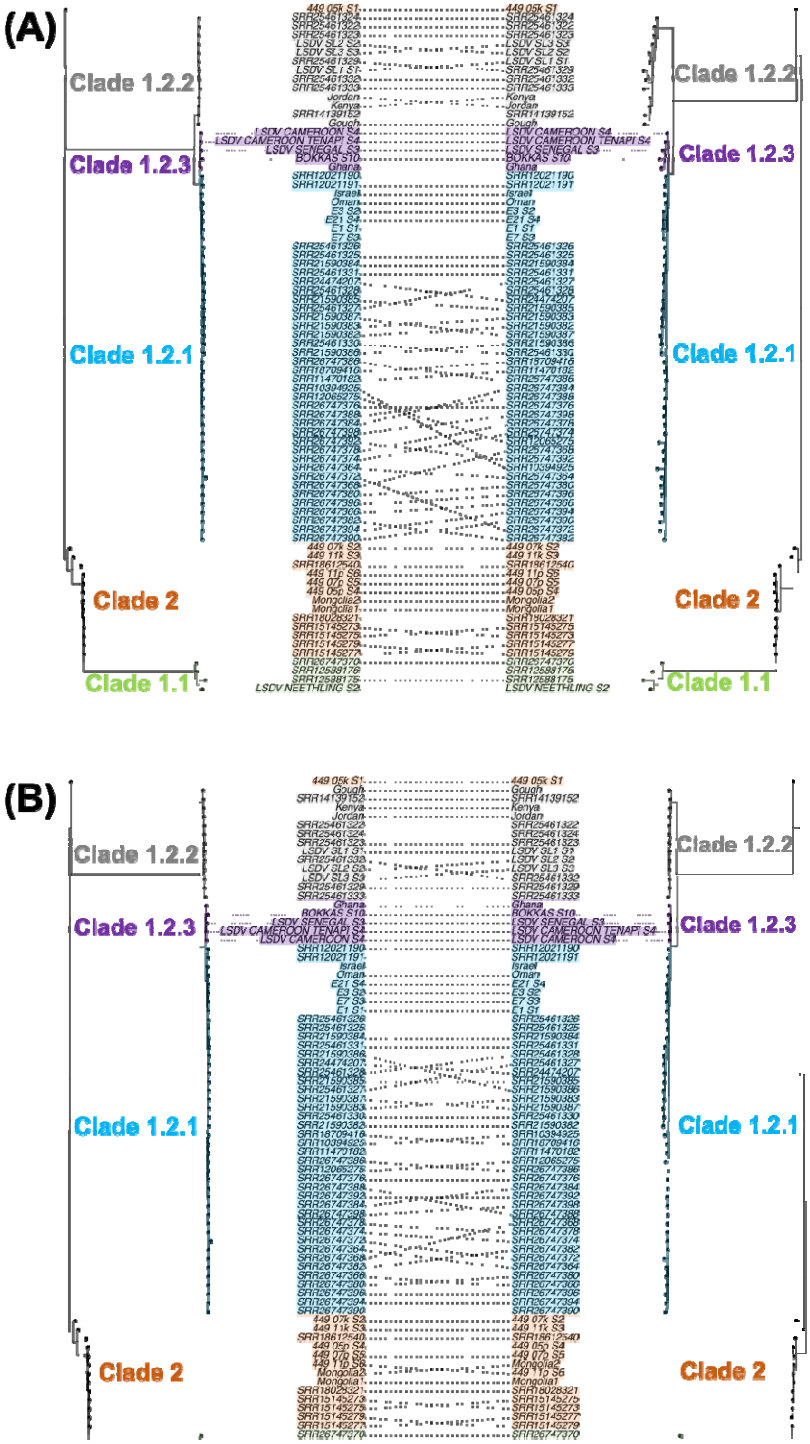


Figure 6. Co-phylogenies of genome-wide SNPs inferred using (left) Minimap2 to a linear reference or (right) Giraffe to a (A) three- or (B) six-sample PVG. The estimated clades are shown by shaded areas: Clade 1.1 is in pale green, subclade 1.2.1 in pale cyan, subclade 1.2.2. is in grey, subclade 1.2.3 is in purple, and Clade 2 in pale orange.

Nonsynonymous (N) and synonymous (S) mutation rates from Minimap2 and Giraffe mapping to the one-sample PVG showed only subtle differences. Average N/S ratios were nearly identical (both 0.44), indicating strong consistency between approaches. Giraffe identified 0.3% more synonymous SNPs, 1.4% more nonsynonymous SNPs, and 2.4% more intergenic SNPs (Table 4). Both methods highlighted genes LD008, LD009, LD134, LD140 and LD145 as having the highest SNP rates; these encode proteins associated with host recognition and immunomodulation. For example, the PVG-based approach uniquely identified E183D in LD067 (encoding a host-range protein) in all Clade 1.1 and Clade 2 samples. A stop loss in LD019a (encoding a kelch-like protein) was detected in seven samples using PVG-based methods but linear mapping found it only three samples.

Group	Minimap2	Giraffe_1	Giraffe_3	Giraffe_1&3
Synonymous	1,301	1,305	324	1,523
Nonsynonymous	581	589	418	894
Intergenic	84	86	46	117
Stop loss	3	3	4	5
Stop gain	1	2	32	34
Nonsyn/syn ratio	0.45	0.45	1.29	0.59

Table 4. Number of synonymous, nonsynonymous, intergenic, stop loss and stop gain SNPs detected in LSDV samples from Clades 1.1, 1.2 and 2. Linear reference mapping with Minimap2 (‘Minimap2’). Giraffe mapping to one-sample PVG (‘Giraffe_1’). Giraffe mapping to three-sample PVG (‘Giraffe_3’). Giraffe with merged one- and three-sample PVG (‘Giraffe_1&3’).

Relative to Minimap2 mapping with a linear reference, Giraffe mapping to the merged one- and three-sample PVGs recovered 17% more synonymous SNPs, 54% more nonsynonymous SNPs, and 39% more intergenic SNPs (Table 4). This also highlighted additional potential functional changes. At LD012 (encoding an ankyrin repeat protein), G136E was present in all subclades 1.2.2 and 1.2.3 samples, and an additional change (Q181*) was in subclade 1.2.2. At LD032 (encoding a polyA polymerase large subunit), S65G and K125* was in subclades 1.2.2 and 1.2.3. At LD135 (IFN-alpha/beta-binding protein), Y117*(128,683 bp) was in six isolates from Mongolia.

In total, 956 SNPs differentiated Clades 1.1 and 1.2 (Table 4). Ten genes accounted for 32% of these: LD134 (accounting for 66 SNPs), LD008 (52 SNPs), LD146 (39), LD009 (25), LD152 (23), LD039 (23), LD135 (22), LD049 (20), LD033 (19) and LD071 (18). Clade 2 showed less differentiation, consistent with its recent recombinant background, but carried three nonsynonymous changes absent in either Clade 1.1 or 1.2: N202D at LD008, N290D at LD059, and E96K at LD135. Within Clade 1.2, two nonsynonymous SNPs distinguished subclade 1.2.1 from 1.2.2/1.2.3, and four further SNPs separated 1.2.2 from 1.2.3.

Discussion

In this study, we constructed the first PVG for LSDV and demonstrated that a small, genetically representative PVG captures nearly all known genomic diversity in the species. A PVG built from only three assemblies, one per major lineage, recovered 97% of known SNP-level diversity while keeping graph size minimal. Comparisons with PVGs built from six and 121 genome assemblies (resulting in PVG extensions of 86 bp and 863 bp, respectively) showed only marginal gains in recovered variation, indicating that LSDV population structure provides a strong guide for selecting PVG representatives. The three-sample PVG retained 81% of all SNPs identified by the full 121-sample PVG while reducing total sequence length from 18,213 Kb to 452 Kb, highlighting the computational efficiency gained by using a representative graph. Notably, read mapping to PVGs has been shown to be as computationally efficient as mapping to a linear reference in both humans [1] and herpesviruses [10], and our findings support this for LSDV.

Our comparisons between PVG-based and linear-reference mapping emphasise the limitations of relying on a single reference genome. Mapping with Giraffe to the one-sample PVG performed comparably to mapping against a linear reference and often outperformed it, consistent with previous analyses in other organisms [9]. Importantly, 27% of SNPs detected using PVGs could not be projected onto the linear reference, demonstrating the extent of reference bias and the inability of a single genome to represent the full spectrum of LSDV haplotypes [24]. PVG-based mapping also improved phylogenetic resolution: it recovered variants that discriminated subclades 1.2.1, 1.2.2 and 1.2.3, whereas linear mapping did not. These gains are particularly relevant for detecting recombinant lineages containing combinations of wild-type and vaccine-derived ancestry [26-29], which continue to emerge across the LSDV geographic range.

Together, these results provide a practical PVG design strategy for viruses with clear population structure to help moderate the compute time versus the information yielded (Figure 7). We propose constructing two complementary PVGs: (i) a one-sample PVG representing the dominant clade, and (ii) a lineage-representative PVG containing one high-quality assembly per major lineage. Mutations identified from both graphs can then be merged without requiring large-scale coordinate conversions, enabling sensitive detection of variants that would be overlooked using linear references alone. The number of genomes included in each PVG should be guided by lineage size, genetic uniqueness, and relevant epidemiological context.

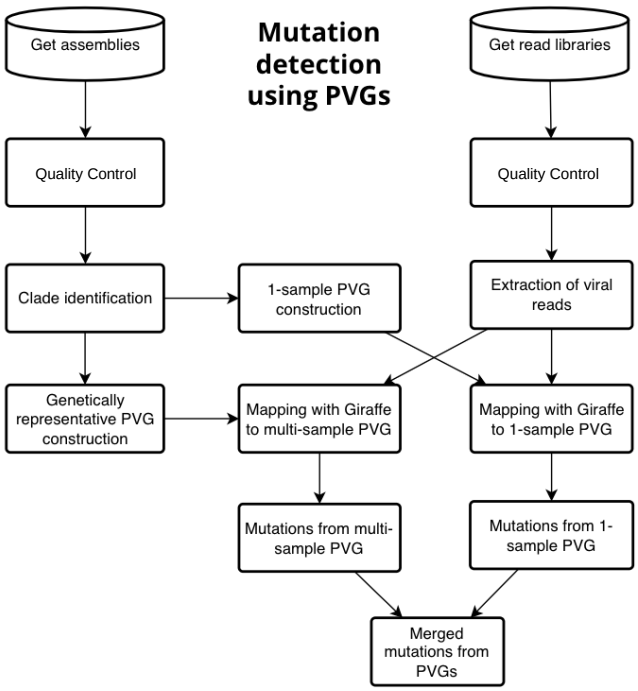


Figure 7. A flow chart for mutation detection using PVGs. Steps start from using genome assemblies to construct a one- and multi-sample PVG. Read libraries can then be mapped to both PVGs to discover mutations by merging the PVG-based outputs.

This study also highlights several limitations. First, WGS read libraries were clearly heterogeneous, whereas amplicon ones were more consistent, and newer datasets generally exhibited higher base quality. The AT-rich LSDV genome (27% GC) [23] likely contributes to variable WGS performance, whereas amplicon sequencing benefited most from PVG-based mapping. Second, although long-read assemblies are ideal for PVG construction because they better resolve structural variants and repetitive regions, some older Nanopore datasets had insufficient base quality for reliable inclusion. Third, while our approach is well-suited to viruses with relatively simple population structures such as LSDV, highly diverse RNA viruses or within-host quasispecies may require more specialised PVG strategies.

Overall, we show that PVGs provide substantial advantages for analysing viral genomic diversity. For LSDV, a small lineage-representative PVG captured nearly all known variation while dramatically reducing computational overhead, and PVG-based mapping, particularly with Giraffe, identified more biologically meaningful variants than linear mapping. Our framework offers a generalisable strategy for building viral PVGs and has direct implications for mixed-sample detection, genomic surveillance, outbreak reconstruction, and identifying recombinant vaccine-derived lineages in LSDV [86] and other DNA viruses.

Supplementary Table Legends

Table S1. The Illumina read libraries' sample names, assigned clade/subclade (Clade), library type (Type), read type (Layout), number of bases (bases), average length (avgLength), size in MB (size_MB), experiment ID (experiment), library name (LibraryName), library selection (LibrarySelection), sequencing machinte (Model), Bioproject ID, biosample ID, sample name and additional information (where relevant). Libraries 449_05k_S1, 449_05p_S4, 449_07k_S2, 449_07p_S5, 449_11k_S3 and 449_11p_S6 were newly sequenced.

Table S2. The quality control metrics for each read library, showing the numbers of reads pre-QC (before_filtering) and post-QC (after_filtering), the number of bases with BQ scores > 30 pre_QC (Q30_bases_before) and post-QC (Q30_bases_after), numbers of failing QC due to low quality (reads_failed_low_quality) or too many N bases (reads_failed_too_many_N) or if they were too short (reads_failed_too_short) or containing adapters (reads_with_adapter_trimmed) or partial adapters (bases_trimmed_due_to_adapters), and the duplicate read numbers (duplication_rate). Paired read libraries were merged so that each sample was represented by one set of read libraries.

Table S3. The number and percentage of reads associated with each read library across the major common name species found (cow, deer, sheep, human). Capripoxvirus reads are allocated under the columns "virus".

Table S4. The reference name (rname), read depth, BQ and MQ across the read libraries organised by their mapping approach. The mapping approaches were Minimap2 to a linear reference, or Giraffe to the 1- or 3- or 6- or all-sample PVGs.

Table S5. Ti/Tv rates for different library types (amplicon and WGS) based on the mapping tool (Minimap2 vs Giraffe), PVG type (1-, 3-, 6-, all-sample or linear) and mutation caller (BCF stands for BCFtools, FB for Freebayes) for all SNPs and invalid ones.

Table S6. Fraction of valid SNPs that could not be lifted over from the sample's coordinates to KX894508 if the allelic options allow. BCF stands for SNPs detected with BCFtools. FB stands for SNPs detected with Freebayes. VG stands for SNPs detected with vg. 3, 6 and ALL stand for mapping approaches using the 3-, 6- and all-sample PVGs, respectively. The mapping approaches using Giraffe are labelled GBWT.

Table S7. The numbers of SNPs discovered using Giraffe mapping to the combined one- and three-sample PVG Giraffe mapping compared to that for Minimap2 to a linear reference across samples. PVG denotes the number of SNPs found by PVG-based mapping. Minimap2 denotes the number of SNPs found by mapping to a linear reference. Total_all denotes the number of SNPs found by either option. Percent_PVG is the percentage of SNPs uniquely found by PVG-based methods, and Percent_M is the percentage of SNPs uniquely found by Minimap2. The samples are sorted by Library_type and then by the number of Minimap2 SNPs.

Table S8. The numbers of SNPs discovered using Giraffe mapping to the combined one- and six-sample PVG Giraffe mapping compared to that for Minimap2 to a linear reference across samples. PVG denotes the number of SNPs found by PVG-based mapping. Minimap2 denotes the number of SNPs found by mapping to a linear reference. Total_all denotes the number of SNPs found by either option. Percent_PVG is the percentage of SNPs uniquely found by PVG-based methods, and Percent_M is the percentage of SNPs uniquely found by Minimap2. The samples are sorted by Library_type and then by the number of Minimap2 SNPs.

Funding information

We had funding support through from UK Research and Innovation (UKRI) Biotechnology and Biological Sciences Research Council (BBSRC) grants BBS/E/PI/230002A, BBS/E/PI/230002B, BBS/E/PI/230002C and BBS/E/PI/23NB0003.

Acknowledgements

We acknowledge the Pirbright Institute's Bioinformatics STP and Non-vesicular Reference Laboratory (NVRL).

Author contributions

TD; Conceptualisation. CT, CW, LD, TD.; Methodology. TD; Investigation. CT; Resources. TD; Data Curation. CT, CW, LD, TD; Writing – Original Draft Preparation. CW, TD; Writing – Review and Editing. CT, CW, LD, TD; Visualisation. TD; Project Administration.

Conflicts of interest

The authors declare no conflicts of interest.

References

1. Eizenga JM, Novak AM, Sibbesen JA, Heumos S, Ghaffaari A, Hickey G, Chang X, Seaman JD, Rounthwaite R, Ebler J, Rautiainen M, Garg S, Paten B, Marschall T, Sirén J, Garrison E. Pangenome Graphs. Annu Rev Genomics Hum Genet. 2020 Aug 31;21:139-162. doi: 10.1146/annurev-genom-120219-080406. Epub 2020 May 26.
2. Hickey G, Heller D, Monlong J, Sibbesen JA, Sirén J, Eizenga JM, et al. Genotyping structural variants in pangenome graphs using the vg toolkit. Genome Biol. 2020;21:35.
3. Boehm E, Kronig I, Neher RA, Eckerle I, Vetter P, Kaiser L; Geneva Centre for Emerging Viral Diseases. Novel SARS-CoV-2 variants: the pandemics within the pandemic. Clin Microbiol Infect. 2021 27(8):1109-1117. doi: 10.1016/j.cmi.2021.05.022
4. Garrison E, Guarracino A. Unbiased pangenome graphs. Bioinformatics. 2023;39(1):btac743. doi: 10.1093/bioinformatics/btac743
5. Rick JA, Brock CD, Lewanski AL, Golcher-Benavides J, Wagner CE. Reference Genome Choice and Filtering Thresholds Jointly Influence Phylogenomic Analyses. Syst Biol. 2024;73(1):76-101. doi: 10.1093/sysbio/syad065
6. Valiente-Mullor C, et al. One is not enough: On the effects of reference genome for the mapping and subsequent analyses of short-reads. PLoS Comput Biol. 2021;17(1):e1008678. doi: 10.1371/journal.pcbi.1008678.
7. Chen NC, Solomon B, Mun T, Iyer S, Langmead B. Reference flow: reducing reference bias using multiple population genomes. Genome Biol. 2021;22(1):8 doi: 10.1186/s13059-020-02229-3.
8. Chen NC, Paulin LF, Sedlazeck FJ, Koren S, Phillippy AM, Langmead B. Improved sequence mapping using a complete reference genome and lift-over. Nat Methods. 2024;21(1):41-49. doi: 10.1038/s41592-023-02069-6
9. Garrison E, Sirén J, Novak AM, Hickey G, Eizenga JM, Dawson ET, Jones W, Garg S, Markello C, Lin MF, Paten B, Durbin R. Variation graph toolkit improves read mapping by representing genetic variation in the reference. Nat Biotechnol. 2018 Oct;36(9):875-879. doi: 10.1038/nbt.4227.
10. Duchen D, Clipman S, Vergara C, Thio CL, Thomas DL, Duggal P, Wojcik GL. A hepatitis B virus (HBV) sequence variation graph improves sequence alignment and sample-specific consensus sequence construction for genetic analysis of HBV. PLoS One. 2024;19(4):e0301069. doi: 10.1371/journal.pone.0301069
11. Novak AM, Garrison E, Paten B. A graph extension of the positional Burrows-Wheeler transform and its applications. Algorithms Mol Biol. 2017;12:18. doi: 10.1186/s13015-017-0109-9.
12. Sirén J, Garrison E, Novak AM, Paten B, Durbin R. Haplotype-aware graph indexes. Bioinformatics. 2020 Jan 15;36(2):400-407. doi: 10.1093/bioinformatics/btz575
13. Sirén J, Monlong J, Chang X, Novak AM, Eizenga JM, Markello C, Sibbesen JA, Hickey G, Chang PC, Carroll A, Gupta N, Gabriel S, Blackwell TW, Ratan A, Taylor KD, Rich SS, Rotter JI, Haussler D, Garrison E, Paten B. Pangenomics enables genotyping of known structural variants in 5202 diverse genomes. Science. 2021 Dec 17;374(6574):abg8871. doi: 10.1126/science.abg8871.
14. Downing T. Approaches to studying virus pangenome variation graphs. ArXiv. 2025. doi: https://arxiv.org/abs/2412.05096
15. Tennakoon C, et al. Panalyze: automated virus pangenome variation graph construction, analysis and annotation. BioRxiv 2025 doi: https://doi.org/10.1101/2025.04.10.646565
16. WOAH. Version adopted by the World Assembly of Delegates of the OIE in 2010, OIE,Paris.: Terrestrial Manual of Lumpy Skin Disease; 2010 [Available from: https://www.woah.org/app/uploads/2021/03/a-reso-2016-public.pdf
17. Casal J, Allepuz A, Miteva A, Pite L, Tabakovsky B, Terzievski D, Alexandrov T, Beltrán-Alcrudo D. Economic cost of lumpy skin disease outbreaks in three Balkan countries: Albania,

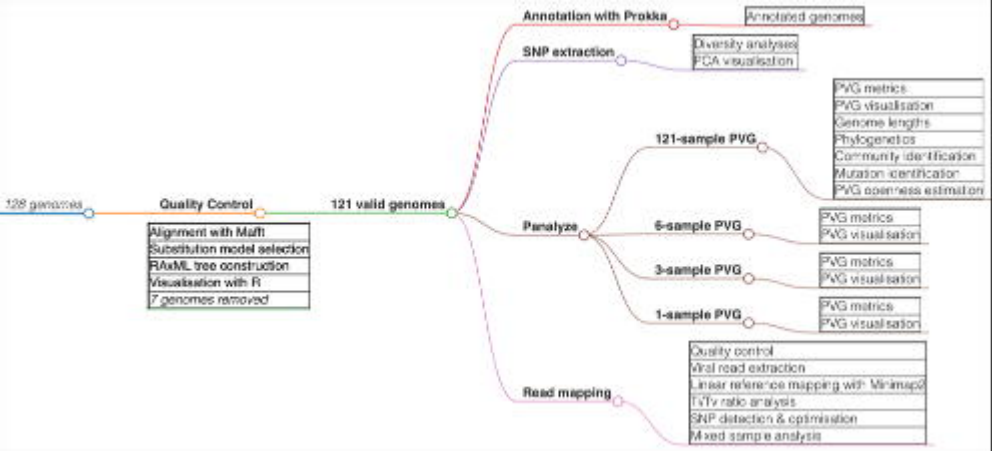
Bulgaria and the Former Yugoslav Republic of Macedonia (2016-2017). Transbound Emerg Dis. 2018 Dec;65(6):1680-1688. doi: 10.1111/tbed.12926.

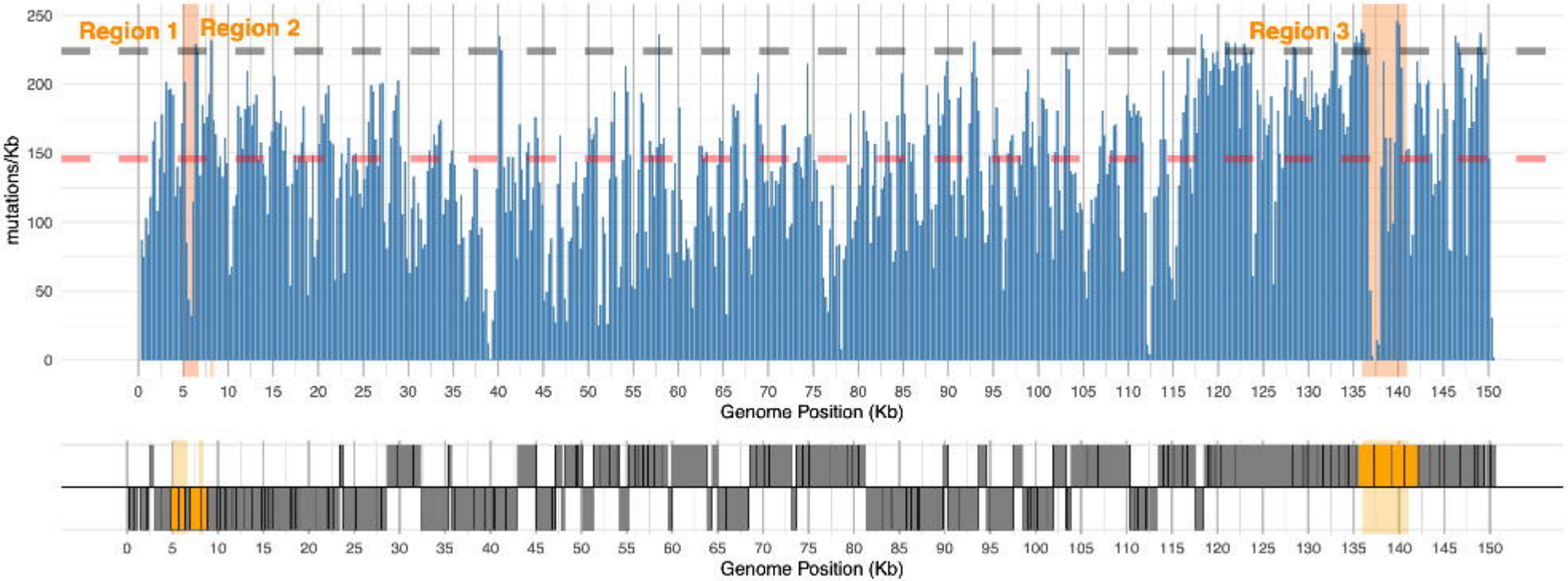
18. Vinitchaikul P, Punyapornwithaya V, Seesupa S, Phuykhamsingha S, Arjkumpa O, Sansamur C, Jarassaeng C. The first study on the impact of lumpy skin disease outbreaks on monthly milk production on dairy farms in Khon Kaen, Thailand. Vet World. 2023 Apr;16(4):687-692. doi: 10.14202/vetworld.2023.687-692.
19. House JA, Wilson TM, Nakashly SE, Karim IA, Ismail I, Danaf NE, et al. The isolation of lumpy skin disease virus and bovine herpesvirus-from cattle in Egypt. J Vet Diagn Invest. 1990;2(2):111–5.
20. Lojkić I, Šimić I, Krešić N, Bedeković T. Complete genome sequence of a lumpy skin disease virus strain isolated from the skin of a vaccinated animal. Genome Announcements. 2018;6(22):e00482–18.
21. Ma J, Yuan Y, Shao J, Sun M, He W, Chen J, Liu Q. Genomic characterization of lumpy skin disease virus in southern China. Transbound Emerg Dis. 2022 Sep;69(5):2788-2799. doi: 10.1111/tbed.14432. Epub 2022 Feb 25. PMID: 34927369.
22. Mathijs E, Vandenbussche F, Nguyen L, Aerts L, Nguyen T, De Leeuw I, Quang M, Nguyen HD, Philips W, Dam TV, Haegeman A, Van Borm S, De Clercq K. Coding-Complete Sequences of Recombinant Lumpy Skin Disease Viruses Collected in 2020 from Four Outbreaks in Northern Vietnam. Microbiol Resour Announc. 2021 Dec 2;10(48):e0089721. doi: 10.1128/MRA.00897-21.
23. Tulman ER, Afonso CL, Lu Z, Zsak L, Kutish GF, Rock DL. Genome of lumpy skin disease virus. J Virol. 2001 75(15):7122-30. doi: 10.1128/JVI.75.15.7122-7130.2001. PMID: 11435593; PMCID: PMC114441
24. Haga IR, et al. Sequencing and Analysis of Lumpy Skin Disease Virus Whole Genomes Reveals a New Viral Subgroup in West and Central Africa. Viruses. 2024;16(4):557. doi: 10.3390/v16040557
25. van Schalkwyk A, Kara P, Heath L. Phylogenomic characterization of historic lumpy skin disease virus isolates from South Africa. Arch Virol. 2022 167(10):2063-2070. doi: 10.1007/s00705-022-05515-6
26. Sprygin A, Van Schalkwyk A, Shumilova I, Nesterov A, Kononova S, Prutnikov P, Byadovskaya O, Kononov A. Full-length genome characterization of a novel recombinant vaccine-like lumpy skin disease virus strain detected during the climatic winter in Russia, 2019. Arch Virol. 2020 165(11):2675-2677. doi: 10.1007/s00705-020-04756-7.
27. Breman FC, Haegeman A, Krešić N, Philips W, De Regge N. Lumpy Skin Disease Virus Genome Sequence Analysis: Putative Spatio-Temporal Epidemiology, Single Gene versus Whole Genome Phylogeny and Genomic Evolution. Viruses. 2023 15(7):1471
28. Van Borm S, et al. Complete genome reconstruction of the global and European regional dispersal history of the lumpy skin disease virus. J Virol. 2023 Nov 30;97(11):e0139423. doi: 10.1128/jvi.01394-23.
29. Sprygin A, Babin Y, Pestova Y, Kononova S, Wallace DB, Van Schalkwyk A, Byadovskaya O, Diev V, Lozovoy D, Kononov A. Analysis and insights into recombination signals in lumpy skin disease virus recovered in the field. PLoS One. 2018 Dec 12;13(12):e0207480. doi: 10.1371/journal.pone.0207480
30. Flannery J, Shih B, Haga IR, Ashby M, Corla A, King S, Freimanis G, Polo N, Tse AC, Brackman CJ, Chan J, Pun P, Ferguson AD, Law A, Lycett S, Batten C, Beard PM. A novel strain of lumpy skin disease virus causes clinical disease in cattle in Hong Kong. Transbound Emerg Dis. 2022 Jul;69(4):e336-e343. doi: 10.1111/tbed.14304. Epub 2021 Sep 16. PMID: 34448540.
31. Li H. Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics. 2018 Sep 15;34(18):3094-3100. doi: 10.1093/bioinformatics/bty191. PMID: 29750242; PMCID: PMC6137996.

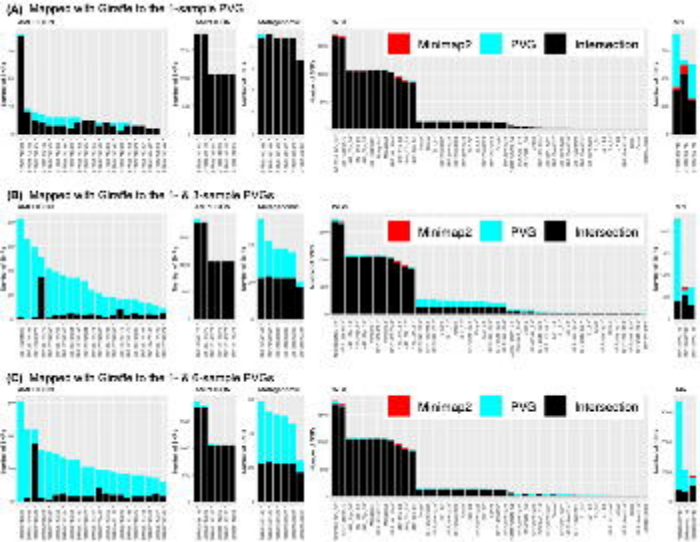
32. Sayers EW, Cavanaugh M, Frisse L, Pruitt KD, Schneider VA, Underwood BA, Yankie L, Karsch-Mizrachi I. GenBank 2025 update. Nucleic Acids Res. 2025 Jan 6;53(D1):D56-D61. doi: 10.1093/nar/gkae1114.
33. Katoh K, Standley DM. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol Biol Evol. 2013;30(4):772–780
34. Kozlov AM, Darriba D, Flouri T, Morel B, Stamatakis A. RAxML-NG: a fast, scalable and user-friendly tool for maximum likelihood phylogenetic inference. Bioinformatics. 2019 35(21):4453–4455.
35. Darriba D, Posada D, Kozlov AM, Stamatakis A, Morel B. ModelTest-NG: A new and scalable tool for the selection of DNA and protein evolutionary models. Mol Biol Evol. 2020;37(1):291–294.
36. R Core Team. R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing; 2024.
37. Paradis E, Schliep K. ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. Bioinformatics. 2019 Feb 1;35(3):526-528. doi: 10.1093/bioinformatics/bty633. PMID: 30016406.
38. Yu G, Smith DK, Zhu H, Guan Y, Lam TT. ggtree: an R package for visualization and annotation of phylogenetic trees with their covariates and other associated data. Methods Ecol Evol. 2017;8:28–36.
39. Schliep KP. phangorn: phylogenetic analysis in R. Bioinformatics. 2011;27(4):592–593.
40. Eddelbuettel D, et al. Rcpp: Seamless R and C++ Integration. J Stat Softw. 2024.
41. Eddelbuettel D, Francois R. RcppArmadillo: Accelerating R with high-performance C++ linear algebra. Comput Stat Data Anal. 2024.
42. Revell LJ. phytools 2.0: an updated R ecosystem for phylogenetic comparative methods (and other things). PeerJ. 2024 Jan 5;12:e16505. doi: 10.7717/peerj.16505. PMID: 38192598; PMCID: PMC10773453.
43. Wang LG, Lam TT, Xu S, Dai Z, Zhou L, Feng T, Guo P, Dunn CW, Jones BR, Bradley T, Zhu H, Guan Y, Jiang Y, Yu G. Treeio: An R Package for Phylogenetic Tree Input and Output with Richly Annotated and Associated Data. Mol Biol Evol. 2020 Feb 1;37(2):599-603. doi: 10.1093/molbev/msz240. PMID: 31633786; PMCID: PMC6993851.
44. Desingu PA, Rubeni TP, Nagarajan K, Sundaresan NR. Sign of APOBEC editing, purifying selection, frameshift, and in-frame nonsense mutations in the microevolution of lumpy skin disease virus. Front Microbiol. 2023 Nov 14;14:1214414. doi: 10.3389/fmicb.2023.1214414. PMID: 38033577; PMCID: PMC10682384.
45. Seemann T. Prokka: rapid prokaryotic genome annotation. Bioinformatics. 2014;30(14):2068–2069
46. Page AJ, et al. SNP-sites: rapid efficient extraction of SNPs from multi-FASTA alignments. Microb Genom. 2016;2:e000056.
47. Cook DE, Andersen EC. VCF-kit: assorted utilities for the variant call format. Bioinformatics. 2017;33(10):1581–1582.
48. Garrison E, et al. vcflib: A C++ library for parsing and manipulating VCF files. 2022.
49. Jombart T, Ahmed I. adegenet 1.3-1: new tools for the analysis of genome-wide SNP data. Bioinformatics. 2011;27:3070–3071.
50. Erik Garrison*, Andrea Guarracino*, Simon Heumos, Flavia Villani, Zhigui Bao, Lorenzo Tattini, Jörg Hagmann, Sebastian Vorbrugg, Santiago Marco-Sola, Christian Kubica, David G. Ashbrook, Kaisa Thorell, Rachel L. Rusholme-Pilcher, Gianni Liti, Emilio Rudbeck, Sven Nahnsen, Zuyu Yang, Mwaniki N. Moses, Franklin L. Nobrega, Yi Wu, Hao Chen, Joep de Ligt, Peter H. Sudmant, Nicole Soranzo, Vincenza Colonna, Robert W. Williams, Pjotr Prins. Building pangenome graphs. Nature Methods (2024). https://doi.org/10.1038/s41592-024-02430-3
51. Guarracino A, Mwaniki N, Marco-Sola S, Garrison E. wfmash: whole-chromosome pairwise alignment using the hierarchical wavefront algorithm. 2021 https://github.com/ekg/wfmash.

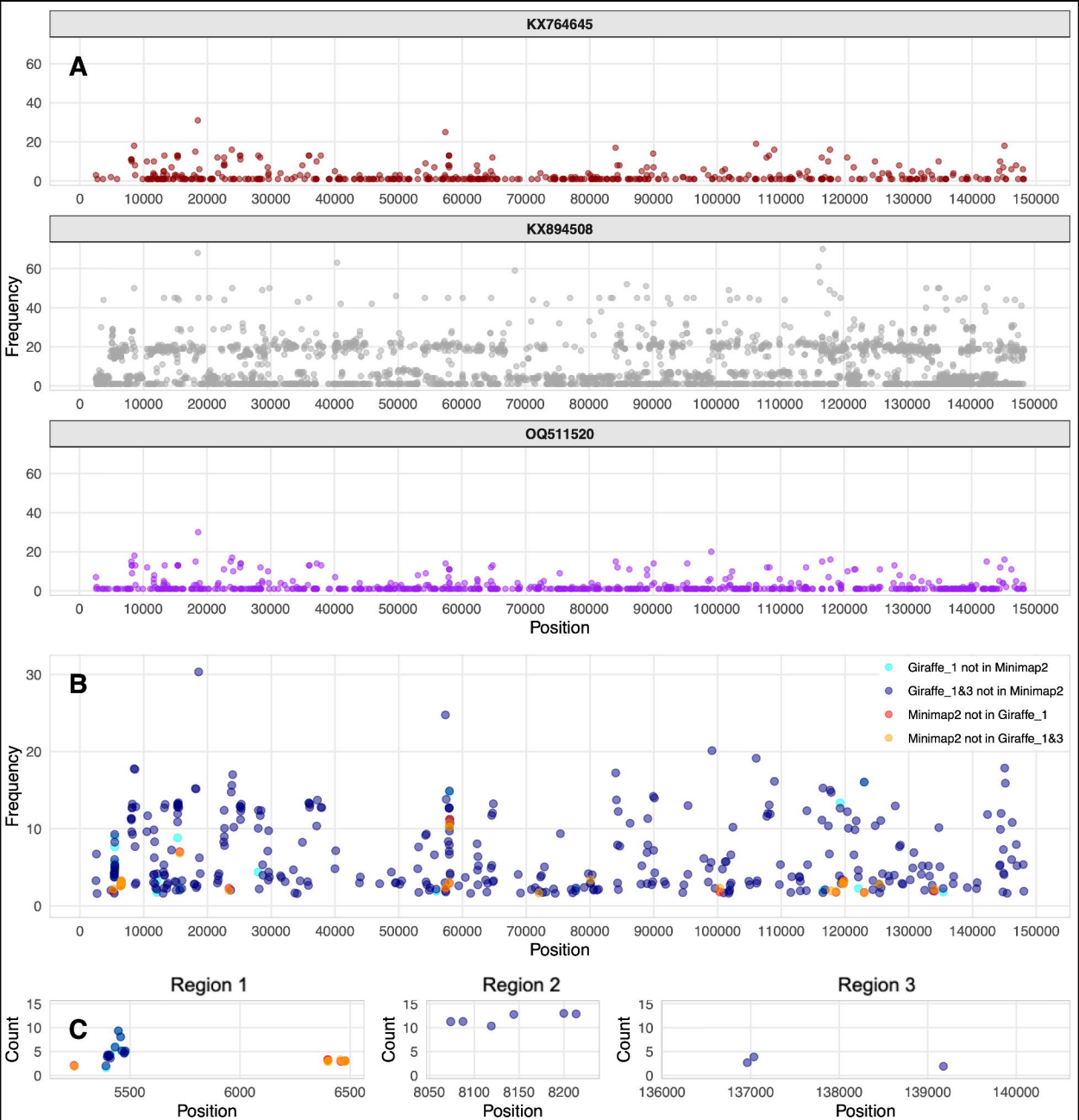
52. Ewels P, Magnusson M, Lundin S, Käller M. MultiQC: summarize analysis results for multiple tools and samples in a single report. Bioinformatics. 2016;32:3047–3048.
53. Danecek P, et al. SAMtools v1.19. Bioinformatics. 2021;37:123–125.
54. Baaijens JA, Bonizzoni P, Boucher C, Della Vedova G, Pirola Y, Rizzi R, Sirén J. Computational graph pangenomics: a tutorial on data structures and their applications. Nat Comput. 2022 Mar;21(1):81-108. doi: 10.1007/s11047-022-09882-6. Epub 2022 Mar 4. PMID: 36969737; PMCID: PMC10038355.
55. Guarracino A, Heumos S, Nahnsen S, Prins P, Garrison E. ODGI: understanding pangenome graphs. Bioinformatics. 2022 Jun 27;38(13):3319-3326. doi: 10.1093/bioinformatics/btac308.
56. Kubica N. gfautil https://crates.io/crates/gfautil
57. Li, H. Gfatools: tools for manipulating sequence graphs in the GFA and rGFA formats. GitHub https://github.com/lh3/gfatools 2024.
58. Formenti G. https://github.com/vgl-hub/gfastats 2025
59. Fischer J. https://github.com/chfi/waragraph 2025
60. Wickham H. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York. 2016. https://ggplot2.tidyverse.org
61. Seppey M, Manni M, Zdobnov EM. BUSCO: Assessing Genome Assembly and Annotation Completeness. Methods Mol Biol. 2019;1962:227-245. doi: 10.1007/978-1-4939-9173-0_14. PMID: 31020564.
62. Pertea G, Pertea M. GFF Utilities: GffRead and GffCompare. F1000Res. 2020 Apr 28;9:ISCB Comm J-304. doi: 10.12688/f1000research.23297.2. PMID: 32489650; PMCID: PMC7222033.
63. Wick RR, Schultz MB, Zobel J, Holt KE. Bandage: interactive visualization of de novo genome assemblies. Bioinformatics. 2015 Oct 15;31(20):3350-2. doi: 10.1093/bioinformatics/btv383. Epub 2015 Jun 22. PMID: 26099265; PMCID: PMC4595904.
64. Garrison EK. fastix. https://github.com/ekg/fastix 2019
65. Csardi G, Nepusz T. The igraph software package for complex network research. InterJournal Complex Systems. 2006;1695.
66. Tille A. art_illumina. 2016. https://manpages.debian.org/testing/art-nextgen-simulation-tools/art_illumina.1.en.html
67. Chen S, Zhou Y, Chen Y, Gu J. fastp: an ultra-fast all-in-one FASTQ preprocessor. Bioinformatics. 2018;34(17):i884–i890.
68. Hannon GJ. FASTX-Toolkit. http://hannonlab.cshl.edu/fastx_toolkit/ 2010
69. Andrews S. FastQC. https://www.bioinformatics.babraham.ac.uk/projects/fastqc/ 2013
70. Nabakooza G, Wagner DD, Momin N, Marine RL, Weldon WC, Oberste MS. Sequence-matching adapter trimmers generate consistent quality and assembly metrics for Illumina sequencing of RNA viruses. BMC Res Notes. 2024 Oct 14;17(1):308. doi: 10.1186/s13104-024-06951-0. PMID: 39402647;
71. Wood DE, Lu J, Langmead B. Improved metagenomic analysis with Kraken 2. Genome Biol. 2019 Nov 28;20(1):257. doi: 10.1186/s13059-019-1891-0. PMID: 31779668; PMCID: PMC6883579
72. Beyer W, Novak AM, Hickey G, Chan J, Tan V, Paten B, Zerbino DR. Sequence tube maps: making graph genomes intuitive to commuters. Bioinformatics. 2019 35(24):5318-5320
73. Garrison E, Marth G. Haplotype-based variant detection from short-read sequencing. arXiv preprint arXiv:1207.3907 2012
74. Lefouili M, Nam K. The evaluation of Bcftools mpileup and GATK HaplotypeCaller for variant calling in non-human species. Sci Rep. 2022 Jul 5;12(1):11331. doi: 10.1038/s41598-022-15563-2. PMID: 35790846; PMCID: PMC9256665.
75. Roder AE, Johnson KEE, Knoll M, Khalfan M, Wang B, Schultz-Cherry S, Banakis S, Kreitman A, Mederos C, Youn JH, Mercado R, Wang W, Chung M, Ruchnewitz D, Samanovic MI, Mulligan MJ, Lässig M, Luksza M, Das S, Gresham D, Ghedin E. Optimized quantification of intra-host viral diversity in SARS-CoV-2 and influenza virus sequence data. mBio. 2023 Aug 31;14(4):e0104623. doi: 10.1128/mbio.01046-23. Epub 2023 Jun 30. PMID: 37389439; PMCID: PMC10470513.

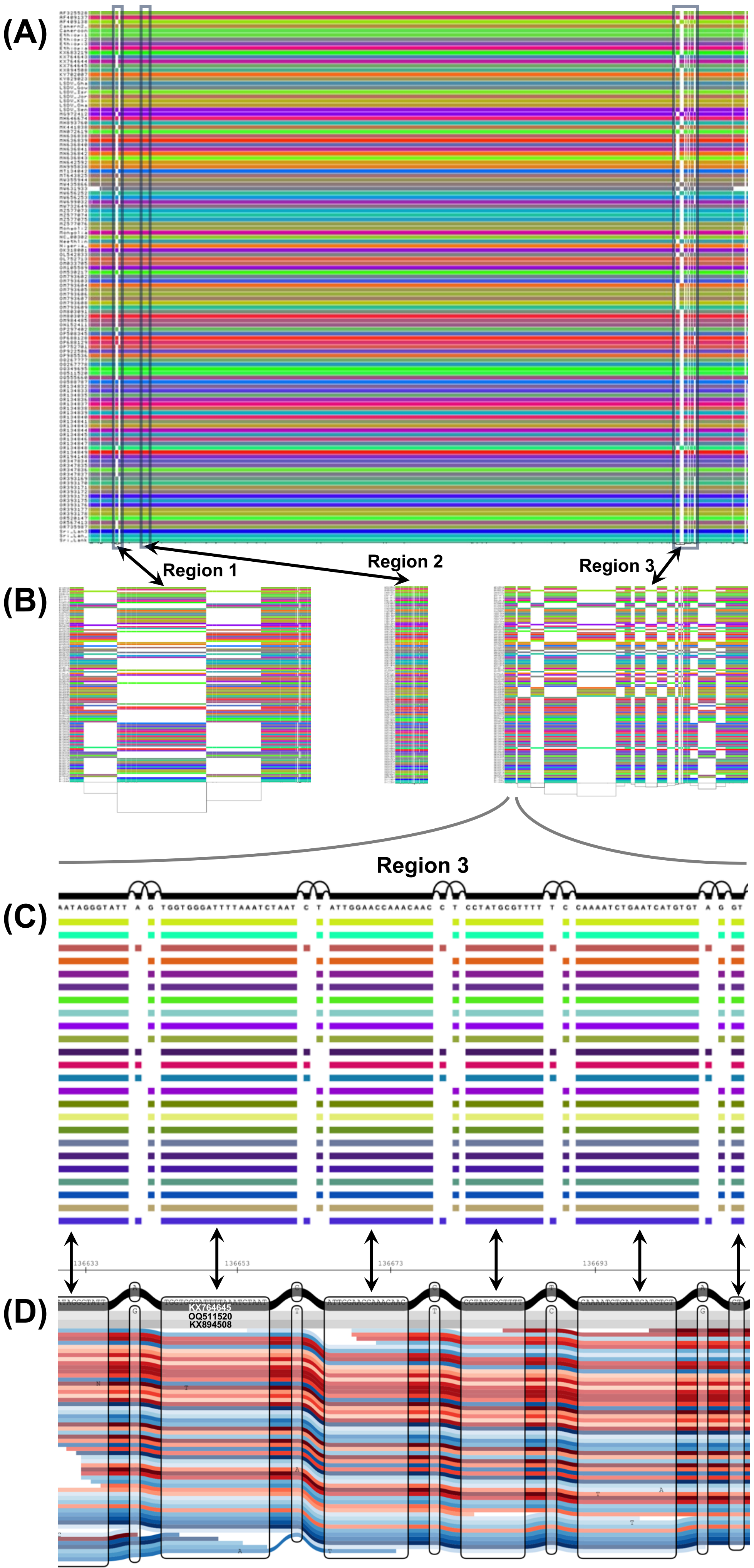
76. Bassano I, Ramachandran VK, Khalifa MS, Lilley CJ, Brown MR, van Aerle R, Denise H, Rowe W, George A, Cairns E, Wierzbicki C, Pickwell ND, Carlile M, Holmes N, Payne A, Loose M, Burke TA, Paterson S, Wade MJ, Grimsley JMS. Evaluation of variant calling algorithms for wastewater-based epidemiology using mixed populations of SARS-CoV-2 variants in synthetic and wastewater samples. Microb Genom. 2023 Apr;9(4):mgen000933. doi: 10.1099/mgen.0.000933. PMID: 37074153; PMCID: PMC10210938.
77. Kamaraj V, Gupta A, Raman K, Narayanan M, Sinha H. GViNC: an innovative framework for genome graph comparison reveals hidden patterns in the genetic diversity of human populations. NAR Genom Bioinform. 2025 Sep 3;7(3):lqaf121. doi: 10.1093/nargab/lqaf121. PMID: 40918067; PMCID: PMC12408910.
78. Thorvaldsdóttir H, Robinson JT, Mesirov JP. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. Brief Bioinform. 2013 14(2):178-92. doi: 10.1093/bib/bbs017
79. Begun DJ, Holloway AK, Stevens K, Hillier LW, Poh YP, Hahn MW, Nista PM, Jones CD, Kern AD, Dewey CN, Pachter L, Myers E, Langley CH. Population genomics: whole-genome analysis of polymorphism and divergence in Drosophila simulans. PLoS Biol. 2007 Nov 6;5(11):e310. doi: 10.1371/journal.pbio.0050310. PMID: 17988176; PMCID: PMC2062478.
80. 1000 Genomes Project Consortium; Abecasis GR, Altshuler D, Auton A, Brooks LD, Durbin RM, Gibbs RA, Hurles ME, McVean GA. A map of human genome variation from population-scale sequencing. Nature. 2010 Oct 28;467(7319):1061-73. doi: 10.1038/nature09534. Erratum in: Nature. 2011 May 26;473(7348):544. Xue, Yali [added]; Cartwright, Reed A [added]; Altshuler, David L [corrected to Altshuler, David]; Kebbel, Andrew [corrected to Keebler, Jonathan]; Koko-Gonzales, Paula [corrected to Kokko-Gonzales, Paula]; Nickerson, Debbie A [corrected to Nickerson, Debo. PMID: 20981092; PMCID: PMC3042601.
81. Guo Y, Li J, Li CI, Long J, Samuels DC, Shyr Y. The effect of strand bias in Illumina short-read sequencing data. BMC Genomics. 2012 Nov 24;13:666. doi: 10.1186/1471-2164-13-666. PMID: 23176052; PMCID: PMC3532123.
82. Wang J, Raskin L, Samuels DC, Shyr Y, Guo Y. Genome measures used for quality control are dependent on gene function and ancestry. Bioinformatics. 2015 Feb 1;31(3):318-23. doi: 10.1093/bioinformatics/btu668. Epub 2014 Oct 8. PMID: 25297068; PMCID: PMC4308666.
83. Hodgkinson A, Eyre-Walker A. Variation in the mutation rate across mammalian genomes. Nat Rev Genet. 2011 Oct 4;12(11):756-66. doi: 10.1038/nrg3098. PMID: 21969038.
84. Pracana R, Priyam A, Levantis I, Nichols RA, Wurm Y. The fire ant social chromosome supergene variant Sb shows low diversity but high divergence from SB. Mol Ecol. 2017 Jun;26(11):2864-2879. doi: 10.1111/mec.14054. Epub 2017 Apr 1. PMID: 28220980; PMCID: PMC5485014.
85. Bergelson et al. Picard Toolkit. Broad Institute, GitHub Repository. http://broadinstitute.github.io/picard/; Broad Institute
86. Biswas S, Noyce RS, Babiuk LA, Lung O, Bulach DM, Bowden TR, Boyle DB, Babiuk S, Evans DH. Extended sequencing of vaccine and wild-type capripoxvirus isolates provides insights into genes modulating virulence and host range. Transbound Emerg Dis. 2020 Jan;67(1):80-97. doi: 10.1111/tbed.13322

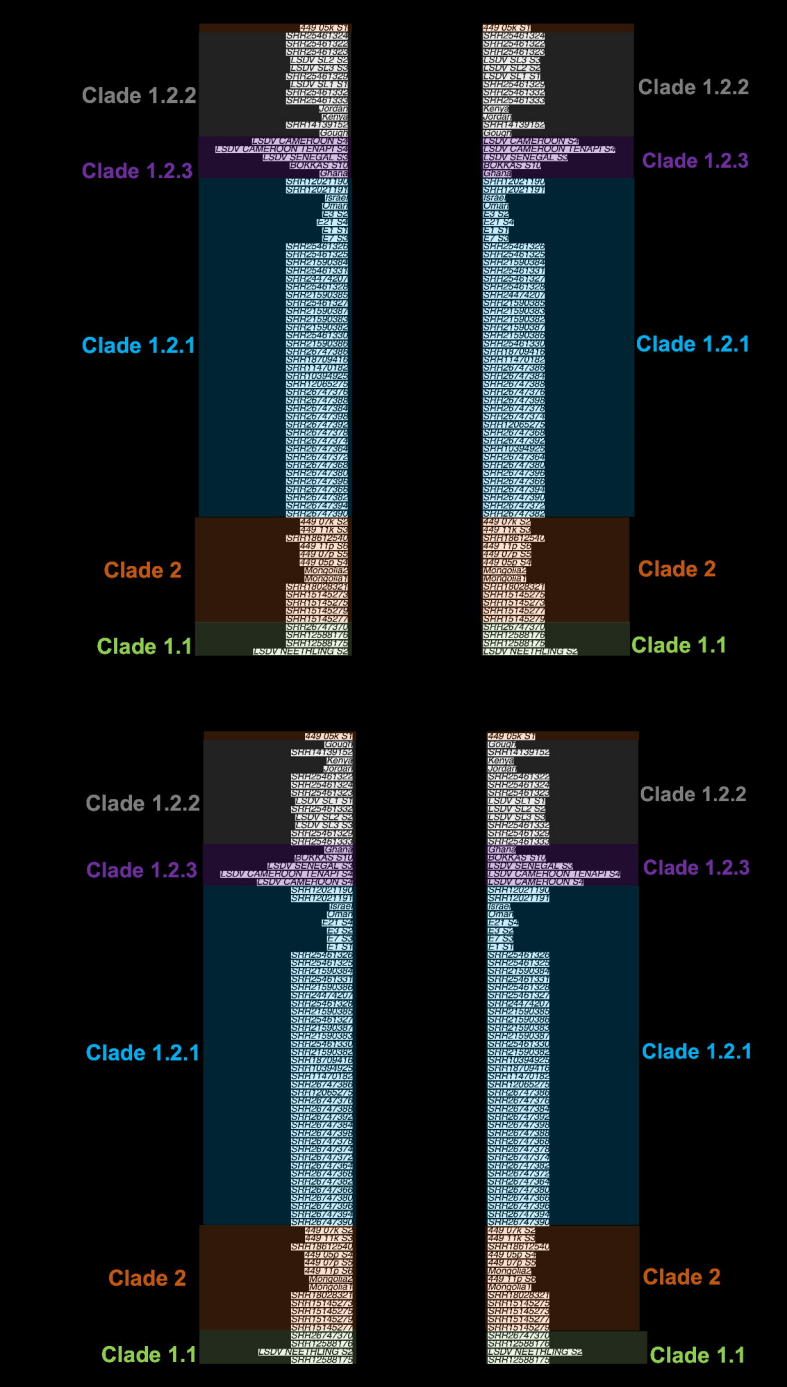

128 genomes
Quality Control
Alignment with Mafft
Substitution model selection
RAxML tree construction
Visualisation with R
7 genomes removed
121 valid genomes
Annotation with Prokka
Annotated genomes
SNP extraction
Diversity analyses
PCA visualisation
Panalyze
121-sample PVG
PVG metrics
PVG visualisation
Genome lengths
Phylogenetics
Community identification
Mutation identification
PVG coverness estimation
6-sample PVG
PVG metrics
PVG visualisation
3-sample PVG
PVG metrics
PVG visualisation
1-sample PVG
PVG metrics
PVG visualisation
Read mapping
Quality control
Viral read extraction
Linear reference mapping with Minimap2
TvTv ratio analysis
SNP detection & optimisation
Mixed sample analysis


Region 1
Region 2
Region 3
mutations/Kb
250
200
150
100
50
0
0 5 10 15 20 25 30 35 40 45 50 55 60 65 70 75 80 85 90 95 100 105 110 115 120 125 130 135 140 145 150
Genome Position (Kb)
0 5 10 15 20 25 30 35 40 45 50 55 60 65 70 75 80 85 90 95 100 105 110 115 120 125 130 135 140 145 150
Genome Position (Kb)


(A) Mappers with Giraffe to the 1-sample PVG
Minimap2
PVG
Intersection
(B) Mappers with Giraffe to the 1- & 3-sample PVGs
Minimap2
PVG
Intersection
(C) Mappers with Giraffe to the 1- & 6-sample PVGs
Minimap2
PVG
Intersection


A
KX764645
KX894508
OQ511520
Frequency
Position
B
Giraffe_1 not in Minimap2
Giraffe_1&3 not in Minimap2
Minimap2 not in Giraffe_1
Minimap2 not in Giraffe_1&3
Frequency
Position
C
Region 1
Region 2
Region 3
Count
Position


(A)
Region 1
Region 2
Region 3
(B)
Region 3
(C)
(D)
136633
136653
136673
136693
KX764645
OQ511520
KX894508


Clade 1.2.2
Clade 1.2.3
Clade 1.2.1
Clade 2
Clade 1.1
Clade 1.2.2
Clade 1.2.3
Clade 1.2.1
Clade 2
Clade 1.1
Clade 1.2.2
Clade 1.2.3
Clade 1.2.1
Clade 2
Clade 1.1
Clade 1.2.2
Clade 1.2.3
Clade 1.2.1
Clade 2
Clade 1.1

Mutation detection using PVGs