Comparative Analysis of Multiple Consensus Genomes of the Same Strain of Marek's Disease Virus Reveals Intrastrain Variation

Alejandro Ortigas-Vasquez[1], Utsav Pandey[2], Daniel Renner[1], Chris Bowen[1], Susan J. Baigent[3], John Dunn[4], Hans Cheng[5], Yongxiu Yao[3], Andrew F. Read[1], Venugopal Nair[3], Dave A. Kennedy[1], Moriah L. Szpara[1,2*]

[1]Departments of Biology and [2]Biochemistry and Molecular Biology, Center for Infectious Disease Dynamics and the Huck Institutes of the Life Sciences, Pennsylvania State University, University Park, Pennsylvania 16802, USA

[3] Viral Oncogenesis Group, The Pirbright Institute, Woking, UK, GU24 0NF

[4] United States Department of Agriculture, Agricultural Research Service, US National Poultry Research Center, Southeast Poultry Research Laboratory, Athens, Georgia 30605, USA

[5] United States Department of Agriculture, Agricultural Research Service, US National Poultry Research Center, Avian Disease and Oncology Laboratory, East Lansing, Michigan 48823, USA

***Corresponding Author**
Dr. Moriah L. Szpara, email: moriah@psu.edu

CRediT (Contributor Roles Taxonomy; https://credit.niso.org/contributor-roles-defined/)

Conceptualization – AOV, MLS, DAK
Data curation – UP, DWR, AOV
Formal analysis – AOV
Funding acquisition – DAK, VN, MLS, AFR
Investigation – UP, CDB, SJB, YY, JD, HC
Methodology – UP, MJJ
Project administration – MLS, DAK
Resources – JD, HC, SJB, YY, VN, DWR, AFR
Software – UP, DWR, AOV
Supervision – MLS, DAK
Validation – CDB, DWR, AOV
Visualization – AOV, MLS, DAK
Writing – AOV
Writing – review & editing – AOV, MLS, DAK

Abstract

Current strategies to understand the molecular basis of Marek's disease virus (MDV) virulence primarily consist of cataloguing divergent nucleotides between strains with different phenotypes. However, each MDV strain is typically represented by a single consensus genome despite the confirmed existence of mixed viral populations. To assess the reliability of single-consensus interstrain genomic comparisons, we obtained two additional consensus genomes of vaccine strain CVI988 (Rispens) and two additional consensus genomes of the very virulent strain Md5 by sequencing viral stocks and cultured field isolates. In conjunction with the published genomes of CVI988 and Md5, this allowed us to perform 3-way comparisons between consensus genomes of the same strain. We found that consensus genomes of CVI988 can vary in as many as 236 positions involving 13 open reading frames (ORFs). In contrast, we found that Md5 genomes varied only in 11 positions involving a single ORF. Phylogenomic analyses showed all three Md5 consensus genomes clustered closely together, while also showing that $CVI988_{GenBank.BAC}$ diverged from $CVI988_{Pirbright.lab}$ and $CVI988_{USDA.PA.field}$. Comparison of CVI988 consensus genomes revealed 19 SNPs in the unique regions of $CVI988_{GenBank.BAC}$ that were not present in either $CVI988_{Pirbright.lab}$ or $CVI988_{USDA.PA.field}$. Finally, we evaluated the genomic heterogeneity of CVI988 and Md5 populations by identifying positions with >2% read support for alternative alleles in two ultra-deeply sequenced samples. We were able to confirm that both populations of CVI988 and Md5 were mixed, exhibiting a total of 29 and 27 high-confidence minor variant positions, respectively. We did not find any evidence of minor variants in the positions corresponding to the 19 SNPs in the unique regions of $CVI988_{GenBank.BAC}$. Taken together, our findings confirm that consensus genomes of the same strain of MDV can vary and suggest that

multiple consensus genomes per strain are needed in order to maximize the accuracy of interstrain genomic comparisons.

Keywords: Intrastrain; Diversity; Population; Genome; MDV; CVI988; Md5; Variation

Introduction

Marek's disease virus (MDV) is an oncogenic, lymphoproliferative and neuropathic alphaherpesvirus affecting chickens (*Mardivirus gallidalpha 2*, Genus *Mardivirus*; Family *Herpesviridae*). It was first reported in 1907 by Hungarian veterinarian József Marek (1). Successful isolation of MDV in culture was only realized six decades later, closely followed by the development of the first vaccine, HPRS-16/Att (2–4). HPRS-16/Att was soon replaced by a better replicating, equally protective vaccine derived from strain FC126 of turkey herpesvirus (5). This vaccine, known as HVT, became the first widely used commercial vaccine for Marek's disease (MD) (6). HVT was successful in reducing MD incidence in commercial flocks until the late 1970s, when reports of strains with increased virulence that were able to bypass vaccine-induced immunity started to emerge across the USA (7, 8). A bivalent vaccine consisting of HVT with MDV serotype 2 strain SB-1 was introduced in response. However, a similar shift in virulence occurred in just over a decade (9, 10). Outside the USA, a vaccine known as CVI988 had been in use since the mid 1970s (11). Developed in the Netherlands by a group of veterinary scientists under Dr. Bart Rispens, CVI988 offered the best protection against the emerging variants, and it quickly became the "gold standard" vaccine around the world (12). CVI988 is colloquially known today as the "Rispens" vaccine for MDV. Though CVI988 has mostly kept MD losses at bay for the past 30 years, there are now reports of strains potentially able to bypass

CVI988-conferred immunity, and the most recent estimate places MDV-associated losses in the poultry industry at around US $1-2 billion per annum (13–16).

Comparative genomics has been a driving force in efforts to understand the molecular basis of MDV-induced virulence ever since the first viral genomes were sequenced over two decades ago (17, 18). The current strategy involves cataloging variable nucleotide positions between the consensus genomes of different MDV strains. Comparisons are made between strains grouped on the basis of their ability to break through vaccine protection, or "pathotype," with the hope that this will help to identify the genes and polymorphisms associated with virulence (19, 20). There are defined pathotypes (i.e., mild/m, virulent/v, very virulent/vv, very virulent +/vv+), with CVI988 being classified as mild and Md5 as very virulent (21). However, despite MDV strains being known to exist as mixed populations, the degree to which the published consensus genomes represent the average for each strain has never been assessed (22–24). In addition, all available consensus genomes for MDV are low-coverage genomes that only account for the major variants (≥50%) present in the original viral population from which they were sequenced. The lack of information on minor variants (≤50%) can be problematic, due to previous reports suggesting that the diversity within a given viral population contributes to the overall pathotype of a strain (25–28). Although the extent of this contribution is still unknown, this nonetheless suggests that failing to account for minor variants can negatively impact our ability to correlate sequencing data to pathotype (21). Finally, many of the available reference genomes underwent cloning steps prior to sequencing. This includes the only available CVI988 consensus genome, which was sequenced from a bacterial artificial chromosome (BAC) copy of the viral genome, followed by plasmid-based cloning thereof (29). The Md5 consensus genome was obtained from

plasmid-based cloning of viral DNA amplified in cell culture (18). The use of BAC and plasmid cloning approaches generates an artificially homogenous population of viral genomes, further moving away from the real-world biology of herpesvirus strains. In addition, by randomly selecting one viral haplotype from a population, these methods inherently reduce the chances of obtaining a representative consensus genome.

To address these issues, it is necessary to test the reliability of available MDV consensus genomes and assess the genomic heterogeneity of MDV populations using next-generation sequencing (NGS) approaches. Whole-genome resequencing (WGR) is a simple yet powerful approach that consists of comparing newly sequenced individuals against a reference (30). Here, we have adapted this strategy for viruses by sequencing multiple populations of the same strain and determining whether these yield a consistent consensus genome, under the assumption that large shifts in genomic variant frequencies will manifest as consensus-level differences. We independently sequenced two populations of the vaccine strain CVI988 and two populations of the very virulent prototype strain Md5, from a combination of viral stocks and cultured field isolates. The newly-sequenced genomes were compared with previously published consensus genomes using whole-genome pairwise alignments. Using this approach, we were able to detect differences between consensus genomes of each strain. Notably, we found 19 single-nucleotide polymorphisms (SNPs) in the unique genomic regions of $CVI988_{GenBank.BAC}$ that were not present in $CVI988_{Pirbright.lab}$ and $CVI988_{USDA.PA.field}$. Next, we assessed the genomic heterogeneity of CVI988 and Md5 by analyzing ultra-deep Illumina sequencing data (>1,000× depth) and identified positions where at least 2% of the reads supported an alternative allele (31). This method enabled detection of minor allele frequency (MAF) variants and allowed us to confirm

that the ultra-deep sequenced populations of CVI988 and Md5 were mixed. These findings highlight the need to sequence multiple consensus genomes of each MDV strain in order to ensure the accuracy of interstrain genomic comparisons and enhance our ability to link viral genotypes to virulence phenotypes.

Methods

Viral culture and DNA isolation of USDA CVI988 and Md5

The virulence phenotypes and targeted (amplicon-based) sequencing of USDA CVI988 and Md5 have been previously described by Dunn and colleagues (21). The sample we refer to as "$CVI988_{USDA.PA.field}$" was collected as a field sample in 2010 in Pennsylvania (MDV 709B_v_2010_PA), and "$Md5_{USDA.lab}$" was collected in Maryland in 1977 (MDV Md5_vv_1977_MD) (21). Chicken embryo fibroblasts (CEFs) were used to culture these virus stocks. Cultures were maintained in 1:1 mixture of Leibovitz's L-15 and McCoy's 5A (LM) media supplemented with fetal bovine serum (FBS), 200 U/ ml penicillin, 20 µg/ml streptomycin, and 2 µg/ml amphotericin B in a 37 °C, 5% CO_2 incubator. Cells were plated with 4% FBS LM media and maintained in 1% FBS LM media. For storage as viral stocks, infected cells were suspended in freezing media composed of 10% DMSO, 45% FBS, and 45% LM media and kept in liquid nitrogen.

DNA library preparation and sequencing of USDA CVI988 and Md5

DNA libraries for USDA CVI988 and Md5 were prepared for next generation sequencing using an Illumina Miseq platform as previously described (32–34). Briefly, genetic material, including

viral nucleocapsid DNA, was quantified by Qubit (Invitrogen, CA) and by a virus specific qPCR. Total DNA was then acoustically sheared using a Covaris M220, with settings at 60 s duration, peak power 50, 10% duty cycle, at 4 °C. DNA was then processed through the Illumina TruSeq Nano DNA prep kit, using the manufactures recommendations, and further quantified by Qubit (Invitrogen, CA), Bioanalyzer (Agilent), and library specific qPCR (KAPA biosystems). Libraries were then diluted to 17pM per manufacturer's recommendation and sequenced using version 3 paired-end (2x300bp length) chemistry.

Viral culture and DNA isolation of Pirbright CVI988 and Md5

The CVI988 stock we refer to as "$CVI988_{Pirbright.lab}$" was obtained from the Pirbright Institute and prepared from commercial CVI988 vaccine, sourced in the UK (Poulvac Marek CVI vaccine; Zoetis), following 2 passages in primary chicken embryo fibroblast cells (CEF). DNA was prepared from approximately $5x10^6$ cells using the DNeasy-96 kit (Qiagen, Hilden, Germany), according to the manufacturer's instructions, and eluted in DNase-free water.

The Md5 stock we refer to as "$Md5_{Pirbright.lab}$" was also obtained from the Pirbright Institute and derived from a 7th duck embryo fibroblast passage stock kindly provided by Dr. A. M. Fadly (Avian Disease and Oncology Laboratory, USA). To amplify this stock, 5-day-old Rhode Island Red chickens were inoculated intra-abdominally with 1000 plaque forming units of virus. Lymphocytes isolated from spleens harvested at 14 days post-infection were cultured with CEF, the cell-associated virus was harvested at 7 days when cytopathic effect was clearly visible, and then further passed in CEF to produce virus stocks. DNA was prepared from approximately $5x10^6$ cells of the 9th passage CEF stock using the DNeasy-96 kit, and eluted in DNase-free water.

DNA library preparation and sequencing of Pirbright CVI988 and Md5

DNA isolated from each MDV culture was quantified for total DNA by Qubit (Invitrogen) and the amount of either Md5 DNA or CVI988 DNA was confirmed by a previously described qPCR assay (35). Samples were then acoustically sheared using Covaris and high-throughput, deep sequencing DNA library preparations were produced using the Kapa HyperPrep Library Kit (Roche) as per manufacturer's protocol, with 14 PCR cycles. Then, samples were processed through a previously described oligo-enrichment procedure, where MDV-specific probes (36) (myBaits, Arbor Biosciences) were used with the myBaits Target Capture Kit (Arbor Biosciences) to enrich for MDV DNA (37). An additional PCR amplification (14 cycles) yielded libraries which were assessed for quality by Qubit, as well as a KAPA-specific qPCR, which allowed for accurate dilutions of libraries to 4 pM before sequencing. Finally, libraries were sequenced using v3 chemistry 300 bp $\times$ 300 bp paired-end sequencing on an Illumina MiSeq, as per the manufacturer's instruction.

Processing of sequencing reads and genome assembly

Viral reads were identified by aligning raw reads in FASTQ files against a local database of 97 MDV genomes using BLASTN with an E-value cutoff of $1x10^{-2}$ (see Supplementary Table 5 for strain names and accession numbers). The selected reads were then subjected to the quality control and preprocessing step (Step 1) of our published viral genome assembly (VirGA) workflow (38). These properly paired reads were then used for *de novo* assembly using metaSPAdes v3.14.0 (39). The resulting file containing the metaSPAdes scaffolds in FASTA format served as input for the remaining steps of VirGA (Steps 3-4), which include genome

linearization, annotation and post-assembly quality assessments. For the reference-guided contig-ordering step, we used a trimmed version (TRL and TRS regions removed) of the published consensus genome of each strain (**Table 1**). New consensus genomes were also trimmed for downstream analyses. Tandem repeats were identified in trimmed consensus genomes using Tandem Repeats Finder (TRF) v4.09 (40). Manual verification and masking of tandem repeats was conducted using Geneious Prime v2023.0.4. To resolve *meq* at the consensus level, reads were mapped to modified reference genomes possessing different isoforms of *meq*. Local sequencing coverages were then used to determine the consensus-level isoform.

Pairwise identity and genetic distance comparisons

Trimmed consensus genomes with masked tandem repeats were aligned using MAFFT v7.505 with default settings (41). The resulting alignment file in clustal format was imported into Geneious to visualize variable positions (SNPs and Indels) between the genomes. Published MDV consensus genomes sequenced from field samples or cultured stocks with PubMed Identifiers (n = 33) and all consensus genomes used for comparative analyses in the present study (n = 6) were aligned using MAFFT (see **Supplementary Table 5** for details). A neighbor-joining (NJ) tree was constructed from the multi-genome alignment file in clustal format using the Geneious Tree Builder tool with a Tamura-Nei genetic distance model and no outgroup.

Identification of minor variant positions

A custom script was used to identify positions with at least 2% read support for an alternative allele – i.e., a minor variant – in genomes with >1,000x coverage. Positions in tandem repeats and homopolymers were excluded. All remaining minor variant positions were verified through manual inspection of BAM alignment files using the Integrative Genomics Viewer (IGV) v2.12.3 (42).

Nucleotide sequence accession numbers

GenBank accession numbers for the four new and two prior viral genomes are listed in Table 1.

Results

Resequencing viral stocks and field isolates of CVI988 and Md5 using Illumina whole-genome sequencing

To enable intrastrain whole-genome comparisons, we first obtained two additional consensus genomes of both vaccine strain CVI988 (mild pathotype) and prototype very virulent strain Md5 (vv pathotype) (**Table 1**). $Md5_{USDA.lab}$ and $Md5_{Pirbright.lab}$ were sequenced from standard lab stocks (i.e., a virus population initiated from a prior viral stock and expanded in host cells). $CVI988_{Pirbright.lab}$ was likewise derived from lab stock. In contrast, $CVI988_{USDA.PA.field}$ was isolated from a sample collected at a Pennsylvania farm, before isolation and expansion in host cells in the lab. Targeted, amplicon-based sequencing of portions of the $Md5_{USDA.lab}$ (Md5_vv_1977_MD) and $CVI988_{USDA.PA.field}$ (709B_v_2010_PA) genomes has been previously reported (21). All viral stocks were propagated similarly for viral DNA isolation and sequencing

library preparation (see Methods for details). $CVI988_{Pirbright.lab}$ and $Md5_{Pirbright.lab}$ were additionally enriched using published methods for oligonucleotide-bait based enrichment in solution (36), which can enable a greater sequencing depth by enriching viral sequences from the overall milieu of DNA. Viral consensus genomes for each stock were determined using a previously published combination of *de novo* and reference-guided assembly methods (**Table 1**; see Methods for details).

The MDV genome has a double-stranded DNA (dsDNA) structure, with a full length of approximately 175-180 kb. The genomic layout is split into two large unique regions known as Unique Long (UL) and Unique Short (US), each flanked by inverted repeat sequences (IRL, IRS) and terminal repeat sequences (TRL, TRS) (**Figure 1A**, center). The α-type packaging sequences are located at the genomic termini and at the IRL/IRS junction. None of the new genomes contained any truncated or missing proteins. The average G+C content of all genomes was 44.0%. The average sequencing coverage of Illumina sequencing data exceeded 100× for all viral genomes and it exceeded 1,000× for the two genomes subjected to oligonucleotide-bait based enrichment (**Table 1**).

Pairwise alignment of same-strain consensus genomes reveals greater intrastrain variation in CVI988 than Md5

For each strain, trimmed consensus genomes with masked tandem repeats were aligned and visualized. Consensus-level differences corresponding to SNPs, insertions and deletions (Indels) were identified and verified through manual inspection of sequence-read alignment files (i.e., BAM files). A total of 236 differences were detected in the three consensus genomes of vaccine strain CVI988, impacting a total of 13 ORFs. These include the genes UL22 / gH (MDV034),

UL49 /VP22 (MDV062), MDV074, MDV075, Meq /RLORF7 (MDV076 / MDV005), vIL-8 / RLORF2 (MDV078 / MDV003), MDV080, MDV085, MDV086, US2 (MDV091), US3 (MDV092), SORF4 / S4 (MDV093) and US6 / gD (MDV094) (**Figure 1A, Supplementary Table 1**). Outside of regions associated with Indels and repetitive elements, $CVI988_{Pirbright.lab}$ and $CVI988_{USDA.PA.field}$ were found to be nearly identical (>99.99% identity), being distinguished only by a single synonymous SNP in UL22. In contrast, $CVI988_{GenBank.BAC}$ possessed 23 unique SNPs and was the only CVI988 genome to exhibit variation in US. On the other hand, the three consensus genomes of very virulent strain Md5 showed a total of 11 differences, with only a single gene, MDV075, being impacted (**Supplementary Table 2**).

The *meq* oncogene contributes significantly to overall intrastrain variation in consensus genomes of CVI988

Out of the 236 differences between CVI988 consensus genomes (**Supplementary Table 1**), 214 were found in IRL (n = 209 bp) and IRS (n = 5 bp). A large portion of these differences (n = 177 bp) corresponded to a 59 amino acid (AA) insertion in the *meq* oncogene, which was found in the $CVI988_{GenBank.BAC}$ and $CVI988_{USDA.PA.field}$ consensus genomes but absent in the $CVI988_{Pirbright.lab}$ consensus genome (**Figure 2**). For *meq*, previous studies have suggested the existence of up to three isoforms: Meq, Long Meq (L-meq) and Short Meq (S-meq) (43–46). Meq is the standard isoform and is 339-AA in length. L-meq is a longer version of Meq that is characterized by a 59-AA insertion in the transactivation domain (TAD), which is often accompanied by a 3-bp CCA deletion immediately upstream from the insertion site. S-meq is a shorter version of Meq that exhibits a 41-AA deletion in the TAD. The CCA deletion that is typically associated with the presence of the L-meq isoform was confirmed to exist as a

subpopulation of considerable size (~ 40%) in the UK CVI988 population, but it did not appear to be linked to the presence of the 177-bp insertion (47). All three Md5 genomes exhibited the standard isoform of *meq* (**Figure 2**), again indicating less diversity when compared to CVI988.

Increased intrastrain variation in structural repeat regions is associated with greater homopolymer instability

Outside of differences associated with *meq* isoforms, 28 of the 37 remaining differences found in IRS and IRL of CVI988 consensus genomes were associated with variation in homopolymer lengths (**Supplementary Table 1**). By contrast, out of the 22 differences found across the unique regions of CVI988 consensus genomes, 20 were ORF-associated SNPs, with an even distribution between synonymous (n = 10) and non-synonymous (n = 10). In Md5 consensus genomes, all 11 differences were found in either IRL (n = 6) or IRS (n = 5) and were all exclusively associated with variation in homopolymer lengths (**Supplementary Table 2**).

Intrastrain variation in unique regions is associated with 19 SNPs in the published CVI988 consensus genome

Out of the 20 ORF-associated SNPs found in unique regions of CVI988 consensus genomes, 19 belonged to $CVI988_{GenBank.BAC}$, including 3 SNPs in UL and 16 SNPs in US (**Figure 1B**). Phylogenomic analyses confirmed divergence of $CVI988_{GenBank.BAC}$ from $CVI988_{Pirbright.lab}$ and $CVI988_{USDA.PA.field}$ (**Figure 3**). Additionally, none of the 3 SNPs found in UL of $CVI988_{GenBank.BAC}$ were present in any of the other 38 MDV genomes included in this study. In contrast, all 16 SNPs in US of $CVI988_{GenBank.BAC}$ were found in subsets across 6 different strains (EU-1, J-1, 814, GX0101, ATE2539 and Kgs-c1). However, no other strain apart from

CVI988$_{GenBank.BAC}$ individually possessed all 16 SNPs in the Unique Short (US) region. The nucleotides found across all 19 positions in CVI988$_{Pirbright.lab}$ and CVI988$_{USDA.PA.field}$ were also found in the three Md5 genomes (**Figure 1B**) and in 27 of the 33 additional MDV genomes used for phylogenomic analyses.

Ultra-deep sequencing reveals presence of minor variants in CVI988 and Md5 populations

In addition to consensus-level comparisons, we performed minor variant analysis on genomes with a sequencing depth exceeding 1,000×. (**Figure 4**, see **Methods** for details). This included CVI988$_{Pirbright.lab}$ (>10,000×) and Md5$_{Pirbright.lab}$ (>8,500×; **Table 1**). A total of 29 minor variant positions were detected in the original CVI988$_{Pirbright.lab}$ population across a combination of coding and regulatory regions. These minor variants impacted a total of 9 ORFs: MDV011, UL13 (MDV025), UL26 (MDV038), MDV069, MDV072, *meq* (MDV076), MDV077, ICP4 (MDV084) and US3 (MDV092) (**Supplementary Table 3).** Similarly, a total of 27 minor variants were detected in the UK Md5 population. These impacted a total of 10 ORFs: UL19 (MDV031), UL41 / VHS (MDV054), UL44 / gC (MDV057), UL48 / VP16 (MDV061), UL52 (MDV066), MDV069, MDV072, MDV078, MDV079 and MDV084 (**Supplementary Table 4**). Twelve of the 14 minor variant positions found across MDV069 and MDV072 were common to both strains. No minor variants were identified in the US region of the Md5$_{Pirbright.lab}$ population. Additionally, we verified that no minor variant positions coincided with the 19 unique SNPs in CVI988$_{GenBank.BAC}$ (**Figure 1B**).

Discussion

The goal of comparative genomics is to provide a link between genotype and phenotype. The accuracy with which these two elements are represented is therefore directly proportional to our ability to make biologically relevant conclusions through the use of comparative genomic methods. DNA viruses have long been considered to be inherently stable, especially when compared to RNA viruses, which are now often described as mutant swarms, or "quasispecies" (48, 49). This had led to the assumption that viral consensus genomes, which represent the majority nucleotide at any given position, can accurately represent the genotype of MDV strains such as vaccine strain CVI988 and the prototype very virulent strain Md5. As demonstrated here and in prior studies, these and other herpesvirus populations can harbor minor variants both in culture and *in vivo* (50–53). Here, by sequencing additional viral stocks and field isolates of these strains, we have sought to evaluate the impact of conducting interstrain comparisons without accounting for the intrinsic genomic heterogeneity of MDV strain populations.

CVI988 is an attenuated and non-oncogenic vaccine strain that has long been known to exist as a mixed population due to undergoing multiple passages *in vitro*, in different laboratories and in multiple vaccine manufacturing companies. There have also been reports suggesting that the effectiveness of CVI988 vaccines can vary depending on the manufacturer, and that even different batches from the same manufacturer can show differences in effectiveness (54, 55). In contrast, Md5 is a virulent field strain of MDV that is currently used as the prototype very virulent (vv) strain in the USDA-ARS-Avian Disease and Oncology Laboratory (ADOL) pathotype classification system (56). As such, the fact that we found 235 differences between consensus genomes of CVI988 while only finding 11 differences between consensus genomes of

Md5 is consistent with their history and real-world biology. Whether the larger number of variant loci in CVI988 reflects this strain's attenuation through serial passage, or if it is a feature of this strain's genetic background, remains to be determined. In addition, the fact that we were able to detect 29 minor variant positions in $CVI988_{Pirbright.lab}$ and 27 minor variant positions in $Md5_{Pirbright.lab}$ confirms that both strains exist as mixed viral populations. Over time, minor variant differences may expand during passage(s) in culture or during *in vivo* infections, leading to consensus level differences (50, 57). Minor variants may thus provide an explanation for the differences observed between consensus genomes of each strain.

Having access to the minor variant profiles of CVI988 and Md5 has the added advantage of providing insight into the potential effects of intrastrain variation in loci of high interest, such as pp38 (MDV073). Currently available qPCR assays to distinguish between CVI988 and virulent MDV strains are based on a single nucleotide polymorphism (SNP) located at position 320 of the pp38 gene (58). Our results show pp38 to be a highly conserved gene between different populations of CVI988 and Md5, with no minor variant positions coinciding with this gene. This not only confirms the robustness of qPCR assays based on the SNP #320 of pp38, but also establishes pp38 as a locus with high levels of within-strain conservation. Overall, our study provides a solid starting point for assessing intrastrain diversity in MDV strains. However, access to additional consensus genomes of CVI988 and Md5 will still be required in order to obtain the complete picture of MDV intrastrain diversity, as well as resequencing other virulent and attenuated MDV strains.

Among our most significant findings was the discovery of 19 SNPs present in the unique regions of $CVI988_{GenBank.BAC}$ but absent in the two new CVI988 consensus genomes sequenced as part of this study. These 19 SNPs were also not present as minor variants in the $CVI988_{Pirbright.lab}$ viral population, where these positions were sequenced at an average depth of >10,000×. CVI988GenBank.BAC was sequenced by Spatz *et al*. in 2007 (29) from a BAC clone of CVI988 at 6× coverage (**Table 1**). This BAC clone was derived by Petherbridge *et al.* in 2003 from a commercial vaccine supplied by Fort Dodge Animal Health (59). As part of the cloning process, a BAC vector was inserted into US2 and the missing region was later reconstructed using PCR and wild-type CVI988 DNA. While explaining the origin of these 19 SNPs in $CVI988_{GenBank.BAC}$ is beyond the scope of our current study, our results strongly suggest that consensus genomes of CVI988 stocks do not typically differ from virulent strains such as Md5, 648a and RB-1B in any of these positions.

The implications of these findings are multiple. Firstly, these 19 SNPs have been reported as differences between CVI988 and virulent strains like Md5, RB-1B, GA and Md11 by many prior studies, starting with the original publication by Spatz *et al.* (60–62). Secondly, the 16 unique SNPs in US of $CVI988_{GenBank.BAC}$ impacting US2, US3, SORF4 and US6 / gD have been proposed as evidence of naturally occurring recombination between CVI988 and virulent strains (63). To date, a total of seven isolates from China have been described as natural recombinants of CVI988, largely on the basis of these 16 SNPs (64–66). The two new consensus genomes of CVI988 generated as part of this study share the same nucleotides in all 16 positions with these isolates, calling into question whether they are in fact natural recombinants of CVI988. An alternative explanation could be that these isolates are examples of freely circulating CVI988

that have reverted to virulence. In such a scenario, the remaining nucleotide differences between these isolates and our two new CVI988 consensus genomes could yield valuable information regarding loci associated with virulence. On the other hand, these isolates could also correspond to vaccine CVI988 reisolated from vaccinated birds. Either way, confirming the status of these seven isolates will require additional consensus genomes of CVI988, ideally from vaccine samples currently being distributed in China. Overall, these findings suggest that a heavy reliance on the published CVI988 consensus genome may have led to a historical overestimation of genomic divergence between CVI988 and virulent strains.

Finally, it is worth mentioning that current Illumina-based approaches do not provide insight into certain regions in MDV genomes where additional intrastrain variation is likely to be found. Specifically, tandem repeats have been suggested to be an important source of diversity in herpesviruses yet remain mostly inaccessible to NGS strategies both at the consensus and sub-consensus levels (52, 67). The use of long read technologies (e.g., Oxford Nanopore and Pacific Biosciences) could potentially help to overcome these limitations and make it possible to reliably assess intrastrain variation at tandem repeats and homopolymers (68, 69). In addition, the use of long reads could enhance our ability to link minor variant positions into proper haplotypes and help us to better understand the real-world biology of MDV strains.

In conclusion, by sequencing multiple viral stocks and field isolates of the vaccine strain CVI988 and the very virulent strain Md5, we have shown that consensus genomes of these strains can differ in multiple positions across all genomic regions. We have also established that these differences are most likely made possible by fluctuations in a mixed population of viral

genomes, the existence of which we have demonstrated using ultra-deep Illumina sequencing. In addition, we have shown that not accounting for the intrinsic genomic heterogeneity of MDV strains may have led to a historical overestimation of divergence between CVI988 and virulent strains, providing an alternative explanation for sequencing data that was previously used to suggest naturally occurring recombination between the two. In doing so, we have demonstrated the value of working with multiple consensus genomes per strain and challenged prevailing notions regarding the genomic stability of DNA viruses.

Acknowledgements

The authors thank members of the Szpara, Kennedy, and Nair labs for helpful feedback and discussion. We thank Matthew J. Jones and Andrew S. Bell for their contributions to early preparation of two samples for this study. This work was supported by NSF-NIH EEID award 1 R01 GM140459 (DK, MS, VN), the Biotechnology and Biological Sciences Research Council (BBSRC) grants BBS/E/I/00007039, BBS/E/PI/23NB0003, BB/CCG2250/1 and BB/V017748/1.

References

1. Marek J. 1907. Multiple Nervenentzündung (Polyneuritis) bei Hühnern. Dtsch Tierärztl Wochenschr 15:417–421.
2. Churchill AE, Biggs PM. 1967. Agent of Marek's disease in tissue culture. Nature 215:528–530.
3. Witter RL, Burgoyne GH, Solomon JJ. 1968. Preliminary studies on cell cultures infected with Marek's disease agent. Avian Dis 12:169–185.
4. Churchill AE, Payne LN, Chubb RC. 1969. Immunization against Marek's disease using a live attenuated virus. Nature 221:744–747.

5. Witter RL, Nazerian K, Purchase HG, Burgoyne GH. 1970. Isolation from turkeys of a cell-associated herpesvirus antigenically related to Marek's disease virus. Am J Vet Res 31:525–538.

6. Okazaki W, Purchase HG, Burmester BR. 1970. Protection against Marek's disease by vaccination with a herpesvirus of turkeys. Avian Dis 14:413–429.

7. Witter RL, Sharma JM, Fadly AM. 1980. Pathogenicity of Variant Marek's Disease Virus Isolants in Vaccinated and Unvaccinated Chickens. Avian Dis 24:210–232.

8. Schat KA, Calnek BW, Fabricant J. 1982. Characterisation of two highly oncogenic strains of Marek's disease virus. Avian Pathol J WVPA 11:593–605.

9. Witter RL. 1982. Protection by attenuated and polyvalent vaccines against highly virulent strains of Marek's disease virus. Avian Pathol J WVPA 11:49–62.

10. Witter RL. 1997. Increased virulence of Marek's disease virus field isolates. Avian Dis 41:149–163.

11. Rispens BH, van Vloten H, Mastenbroek N, Maas HJ, Schat KA. 1972. Control of Marek's disease in the Netherlands. I. Isolation of an avirulent Marek's disease virus (strain CVI 988) and its use in laboratory vaccination trials. Avian Dis 16:108–125.

12. Schat KA. 2016. History of the First-Generation Marek's Disease Vaccines: The Science and Little-Known Facts. Avian Dis 60:715–724.

13. Schumacher D, Tischer BK, Teifke J-P, Wink K, Osterrieder N. 2002. Generation of a permanent cell line that supports efficient growth of Marek's disease virus (MDV) by constitutive expression of MDV glycoprotein E. J Gen Virol 83:1987–1992.

14. Zhang Y, Li Z, Bao K, Lv H, Gao Y, Gao H, Qi X, Cui H, Wang Y, Ren X, Wang X, Liu C. 2015. Pathogenic characteristics of Marek's disease virus field strains prevalent in China and the effectiveness of existing vaccines against them. Vet Microbiol 177:62–68.

15. Sun G-R, Zhang Y-P, Lv H-C, Zhou L-Y, Cui H-Y, Gao Y-L, Qi X, Wang Y-Q, Li K, Gao L, Pan Q, Wang X-M, Liu C-J. 2017. A Chinese Variant Marek's Disease Virus Strain with Divergence between Virulence and Vaccine Resistance. Viruses 9:E71.

16. Davison F, Venugopal N. 2004. Marek's Disease: An Evolving Problem, 1st ed. Academic Press.

17. Lee LF, Wu P, Sui D, Ren D, Kamil J, Kung HJ, Witter RL. 2000. The complete unique long sequence and the overall genomic organization of the GA strain of Marek's disease virus. Proc Natl Acad Sci U S A 97:6091–6096.

18. Tulman ER, Afonso CL, Lu Z, Zsak L, Rock DL, Kutish GF. 2000. The genome of a very virulent Marek's disease virus. J Virol 74:7980–7988.

19. Firth C, Kitchen A, Shapiro B, Suchard MA, Holmes EC, Rambaut A. 2010. Using time-structured data to estimate evolutionary rates of double-stranded DNA viruses. Mol Biol Evol 27:2038–2051.

20. Renzette N, Gibson L, Jensen JD, Kowalik TF. 2014. Human cytomegalovirus intrahost evolution-a new avenue for understanding and controlling herpesvirus infections. Curr Opin Virol 8:109–115.

21. Dunn JR, Black Pyrkosz A, Steep A, Cheng HH. 2019. Identification of Marek's disease virus genes associated with virulence of US strains. J Gen Virol 100:1132–1139.

22. Kuny CV, Szpara ML. 2021. Alphaherpesvirus Genomics: Past, Present and Future. Curr Issues Mol Biol 42:41–80.

23. Renner DW, Szpara ML. 2018. Impacts of Genome-Wide Analyses on Our Understanding of Human Herpesvirus Diversity and Evolution. J Virol 92:e00908-17.

24. Gimeno IM, Cortes AL, Faiz NM, Barbosa T, Villalobos T. 2015. Evaluation of Factors Influencing Efficacy of Vaccine Strain CVI988 Against Marek's Disease in Meat-Type Chickens. Avian Dis 59:400–409.

25. Gimeno IM, Witter RL, Reed WM. 1999. Four distinct neurologic syndromes in Marek's disease: effect of viral strain and pathotype. Avian Dis 43:721–737.

26. Spatz SJ, Volkening JD, Gimeno IM, Heidari M, Witter RL. 2012. Dynamic equilibrium of Marek's disease genomes during in vitro serial passage. Virus Genes 45:526–536.

27. Spatz SJ. 2010. Accumulation of attenuating mutations in varying proportions within a high passage very virulent plus strain of Gallid herpesvirus type 2. Virus Res 149:135–142.

28. Hildebrandt E, Dunn JR, Perumbakkam S, Niikura M, Cheng HH. 2014. Characterizing the molecular basis of attenuation of Marek's disease virus via in vitro serial passage identifies de novo mutations in the helicase-primase subunit gene UL5 and other candidates associated with reduced virulence. J Virol 88:6232–42.

29. Spatz SJ, Petherbridge L, Zhao Y, Nair V. 2007. Comparative full-length sequence analysis of oncogenic and vaccine (Rispens) strains of Marek's disease virus. J Gen Virol 88:1080–1096.

30. Stratton M. 2008. Genome resequencing and genetic variation. Nat Biotechnol 26:65–66.

31. Lauring AS. 2020. Within-Host Viral Diversity: A Window into Viral Evolution. Annu Rev Virol 7:63–81.

32. Bell AS, Kennedy DA, Jones MJ, Cairns CL, Pandey U, Dunn PA, Szpara ML, Read AF. 2019. Molecular epidemiology of Marek's disease virus in central Pennsylvania, USA. Virus Evol 5.

33. Pandey U, Bell AS, Renner DW, Kennedy DA, Shreve JT, Cairns CL, Jones MJ, Dunn PA, Read AF, Szpara ML. 2016. DNA from dust: Comparative genomics of large DNA viruses in field surveillance samples. mSphere 1:e00132-16.

34. Bowen CD, Renner DW, Shreve JT, Tafuri Y, Payne KM, Dix RD, Kinchington PR, Gatherer D, Szpara ML. 2016. Viral forensic genomics reveals the relatedness of classic herpes simplex virus strains KOS, KOS63, and KOS79. Virology 492:179–186.

35. Kennedy DA, Cairns C, Jones MJ, Bell AS, Salathé RM, Baigent SJ, Nair VK, Dunn PA, Read AF. 2017. Industry-Wide Surveillance of Marek's Disease Virus on Commercial Poultry Farms. Avian Dis 61:153–164.

36. Trimpert J, Groenke N, Jenckel M, He S, Kunec D, Szpara ML, Spatz SJ, Osterrieder N, McMahon DP. 2017. A phylogenomic analysis of Marek's disease virus reveals independent paths to virulence in Eurasia and North America. Evol Appl 10:1091–1101.

37. Shipley MM, Rathbun MM, Szpara ML. 2020. Oligonucleotide Enrichment of HSV-1 Genomic DNA from Clinical Specimens for Use in High-Throughput Sequencing. Methods Mol Biol Clifton NJ 2060:199–217.

38. Parsons LR, Tafuri YR, Shreve JT, Bowen CD, Shipley MM, Enquist LW, Szpara ML. 2015. Rapid genome assembly and comparison decode intrastrain variation in human alphaherpesviruses. mBio 6:e02213-14.

39. Nurk S, Meleshko D, Korobeynikov A, Pevzner PA. 2017. metaSPAdes: a new versatile metagenomic assembler. Genome Res 27:824–834.

40. Benson G. 1999. Tandem repeats finder: a program to analyze DNA sequences. Nucleic Acids Res 27:573–580.

41. Katoh K, Misawa K, Kuma K, Miyata T. 2002. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. Nucleic Acids Res 30:3059–3066.

42. Robinson JT, Thorvaldsdóttir H, Winckler W, Guttman M, Lander ES, Getz G, Mesirov JP. 2011. Integrative genomics viewer. Nat Biotechnol 29:24–26.

43. Sato J, Murata S, Yang Z, Kaufer BB, Fujisawa S, Seo H, Maekawa N, Okagawa T, Konnai S, Osterrieder N, Parcells MS, Ohashi K. 2022. Effect of Insertion and Deletion in the Meq Protein Encoded by Highly Oncogenic Marek's Disease Virus on Transactivation Activity and Virulence. Viruses 14:382.

44. Murata S, Yamamoto E, Sakashita N, Maekawa N, Okagawa T, Konnai S, Ohashi K. 2021. Research Note: Characterization of S-Meq containing the deletion in Meq protein's transactivation domain in a Marek's disease virus strain in Japan. Poult Sci 100:101461.

45. Lee SI, Takagi M, Ohashi K, Sugimoto C, Onuma M. 2000. Difference in the meq gene between oncogenic and attenuated strains of Marek's disease virus serotype 1. J Vet Med Sci 62:287–292.

46. Renz KG, Cooke J, Clarke N, Cheetham BF, Hussain Z, Fakhrul Islam AFM, Tannock GA, Walkden-Brown SW. 2012. Pathotyping of Australian isolates of Marek's disease virus and association of pathogenicity with *meq* gene polymorphism. Avian Pathol 41:161–176.

47. Bell AS, Kennedy DA, Jones MJ, Cairns CL, Pandey U, Dunn PA, Szpara ML, Read AF. 2019. Molecular epidemiology of Marek's disease virus in central Pennsylvania, USA. Virus Evol 5:vey042.

48. Domingo E, Perales C. 2019. Viral quasispecies. PLoS Genet 15:e1008271.

49. Singh K, Mehta D, Dumka S, Chauhan AS, Kumar S. 2023. Quasispecies Nature of RNA Viruses: Lessons from the Past. Vaccines 11:308.

50. Trimpert J, Groenke N, Kunec D, Eschke K, He S, McMahon DP, Osterrieder N. 2019. A proofreading-impaired herpesvirus generates populations with quasispecies-like structure. Nat Microbiol https://doi.org/10.1038/s41564-019-0547-x.

51. Shipley MM, Renner DW, Ott M, Bloom DC, Koelle DM, Johnston C, Szpara ML. 2018. Genome-wide surveillance of genital herpes simplex virus type 1 from multiple anatomic sites over time. J Infect Dis 218:595–605.

52. Renner DW, Szpara ML. 2018. The impacts of genome-wide analyses on our understanding of human herpesvirus diversity and evolution. J Virol 92:e00908-17.

53. Loncoman CA, Vaz PK, Coppo MJ, Hartley CA, Morera FJ, Browning GF, Devlin JM. 2017. Natural recombination in alphaherpesviruses: Insights into viral evolution through full genome sequencing and sequence analysis. Infect Genet Evol J Mol Epidemiol Evol Genet Infect Dis 49:174–185.

54. López de Juan Abad BA, Cortes AL, Correa M, Gimeno IM. 2019. Evaluation of Factors That Influence Dose Variability of Marek's Disease Vaccines. Avian Dis 63:591–598.

55. Dambrine G, Labaille J, Boissel É, Dupuy C, Rasschaert D. 2016. Marek's disease herpesvirus: a model of vaccine-dependent adaptation. Virol Montrouge Fr 20:273–286.

56. Witter RL, Calnek BW, Buscaglia C, Gimeno IM, Schat KA. 2005. Classification of Marek's disease viruses according to pathotype: philosophy and methodology. Avian Pathol J WVPA 34:75–90.

57. Kuny CV, Bowen CD, Renner DW, Johnston CM, Szpara ML. 2020. In vitro evolution of herpes simplex virus 1 (HSV-1) reveals selection for syncytia and other minor variants in cell culture. Virus Evol 6:veaa013.

58. Baigent SJ, Nair VK, Le Galludec H. 2016. Real-time PCR for differential quantification of CVI988 vaccine virus and virulent strains of Marek's disease virus. J Virol Methods 233:23–36.

59. Petherbridge L, Howes K, Baigent SJ, Sacco MA, Evans S, Osterrieder N, Nair V. 2003. Replication-competent bacterial artificial chromosomes of Marek's disease virus: novel tools for generation of molecularly defined herpesvirus vaccines. J Virol 77:8712–8718.

60. Spatz SJ, Silva RF. 2007. Sequence determination of variable regions within the genomes of gallid herpesvirus-2 pathotypes. Arch Virol 152:1665–1678.

61. Spatz SJ, Rue CA. 2008. Sequence determination of a mildly virulent strain (CU-2) of Gallid herpesvirus type 2 using 454 pyrosequencing. Virus Genes 36:479–489.

62. Spatz SJ, Zhao Y, Petherbridge L, Smith LP, Baigent SJ, Nair V. 2007. Comparative sequence analysis of a highly oncogenic but horizontal spread-defective clone of Marek's disease virus. Virus Genes 35:753–766.

63. He L, Li J, Zhang Y, Luo J, Cao Y, Xue C. 2018. Phylogenetic and molecular epidemiological studies reveal evidence of recombination among Marek's disease viruses. Virology 516:202–209.

64. He L, Li J, Peng P, Nie J, Luo J, Cao Y, Xue C. 2020. Genomic analysis of a Chinese MDV strain derived from vaccine strain CVI988 through recombination. Infect Genet Evol J Mol Epidemiol Evol Genet Infect Dis 78:104045.

65. Xu H, Krieter AL, Ponnuraj N, Tien YY-T, Kim T, Jarosinski KW. 2022. Coinfection in the host can result in functional complementation between live vaccines and virulent virus. Virulence 13:980–989.

66. Zhang Y, Lan X, Wang Y, Lin Y, Yu Z, Guo R, Li K, Cui H, Qi X, Wang Y, Gao L, Pan Q, Liu A, Gao Y, Wang X, Liu C. 2022. Emerging natural recombinant Marek's disease virus

between vaccine and virulence strains and their pathogenicity. Transbound Emerg Dis 69:e1702–e1709.

67. van Belkum A, Scherer S, van Alphen L, Verbrugh H. 1998. Short-sequence DNA repeats in prokaryotic genomes. Microbiol Mol Biol Rev MMBR 62:275–293.

68. Rhoads A, Au KF. 2015. PacBio Sequencing and Its Applications. Genomics Proteomics Bioinformatics 13:278–289.

69. Wang Y, Zhao Y, Bollas A, Wang Y, Au KF. 2021. Nanopore sequencing technology, bioinformatics and applications. Nat Biotechnol 39:1348–1365.

Tables

Table 1: Sequencing statistics for new and previously published CVI988 and Md5 consensus genomes

MDV Strain	# MDV-specific reads	# Reads used for assembly	Coverage	GenBank Accession	Genome length (bp)	Prior citation
$CVI988_{GenBank.BAC}$	N/A	N/A	6x	DQ530348	178,311	(29)
$CVI988_{USDA.PA.field}$	227,487	138,122	149x	PP032833	176,946	-
$CVI988_{Pirbright.lab}$	11,997,583*	9,730,216	10,055x	PP032835	176,532	-
$Md5_{Genbank}$	N/A	N/A	6x	NC_002229	177,874	(18)
$Md5_{USDA.lab}$	1,030,131	727,124	793x	PP032832	176,724	-
$Md5_{Pirbright.lab}$	10,138,498*	8,296,132	8,530x	PP032834	178,276	-

*The UK viral stock populations were subjected to oligo-bait enrichment to increase the yield of viral DNA (36, 37).

Figure Legends

Figure 1: Comparative analysis of CVI988 and Md5 consensus genomes

A) A total of three trimmed consensus genomes (terminal repeat long (TRL) and short (TRS) removed) were obtained for each strain (CVI988 = Blue, Md5 = Red) and aligned using MAFFT. Consensus-level differences including SNPs and Indels are highlighted (dark lines) based on their relative locations in each respective genome (UL = Unique Long, IRL = Internal Repeat Long, IRS = Internal Repeat Short, US = Unique Short). The total number of differences between genomes of the same strain are indicated on the left, with a breakdown for each individual genome indicated on the right; these loci are all listed in detail in **Supplementary Table 1** and **Supplementary Table 2**. For each strain, ORFs with consensus-level differences in at least one of the three genomes belonging to that strain are highlighted and labeled (with alternative names when possible). The total number of ORFs with consensus-level differences is 13 for CVI988 and 1 for Md5. **B)** ORF-associated SNPs found between consensus genomes of CVI988. Genomic positions are indicated relative to $CVI988_{GenBank.BAC}$, which accounted for 21 out of 23 ORF-associated SNPs. Completely unique SNPs that were not found in any of the other 38 MDV genomes included in the study are highlighted (black borders).

Figure 2: Comparison of the *meq* oncogene in Md5 and CVI988 consensus genomes

Schematic representation of IRL *meq* isoforms in CVI988 (blue) and Md5 (red) consensus genomes. The *meq* oncogene is comprised of the proline/glutamine (Pro/Gln) rich domain, the basic region (BR), the leucine zipper (ZIP) and the transactivation domain (TAD) at the C-

terminal region. Differences in the AA sequence of the two strains are highlighted and labeled (A = Alanine, K = Lysine, S = Serine, E = Glutamic acid). The Meq protein is 339 AA in length. The L-Meq isoform contains a 59-amino acid insertion in the transactivation domain, resulting in a total length of 398 AA.

Figure 3: Neighbor-joining tree of previously published MDV strains and the new consensus genomes of CVI988 and Md5

Consensus genomes of CVI988, Md5 and 33 additional MDV strains from North America, Asia and Europe were aligned using MAFFT (**Supplementary Table 5**). The resulting MSA was used to infer a neighbor-joining tree using Geneious Prime v2023.0.4. CVI988 (blue arrows) and Md5 (red arrows) genomes used for comparative analyses are highlighted.

Figure 4: Minor variant distribution in Pirbright CVI988 and Md5 populations

For each strain (CVI988 = Blue, Md5 = Red), a single viral population was ultra-deep sequenced (>1,000×) to enable identification of minor variants (see **Table 1**). Positions outside of tandem repeats and homopolymers where the minor allele was present in ≥ 2% of the reads were identified as minor variants. Each minor variant was classified depending on its relative location to ORFs and its impact on the resulting amino-acid sequence (when applicable). ORFs associated with minor variants are labeled and highlighted beneath the graph using the same criteria.

Figures

Figure 1: Comparative analysis of CVI988 and Md5 consensus genomes

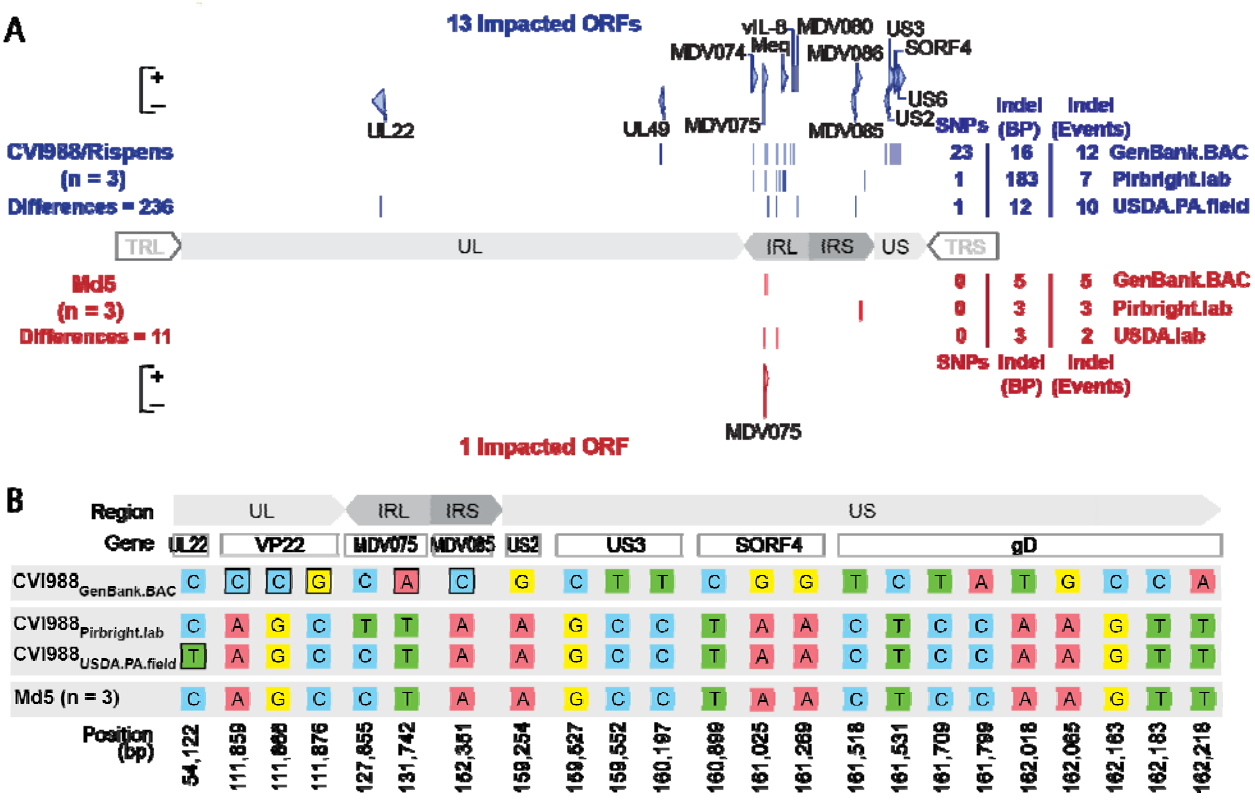


Figure 2: Comparison of the *meq* oncogene in CVI988 and Md5 consensus genomes

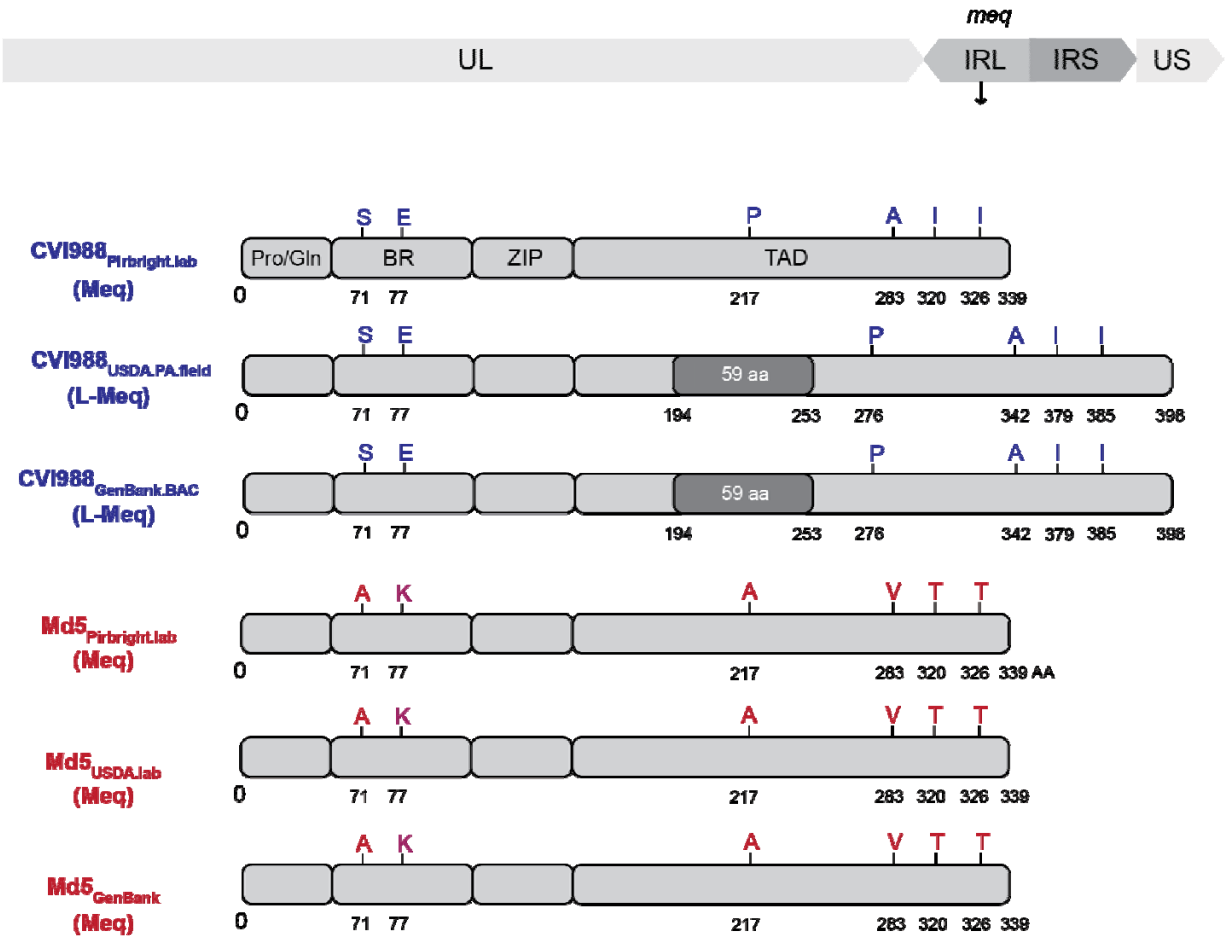


Figure 3: Neighbor-joining tree of previously published MDV strains and consensus genomes of CVI988 and Md5

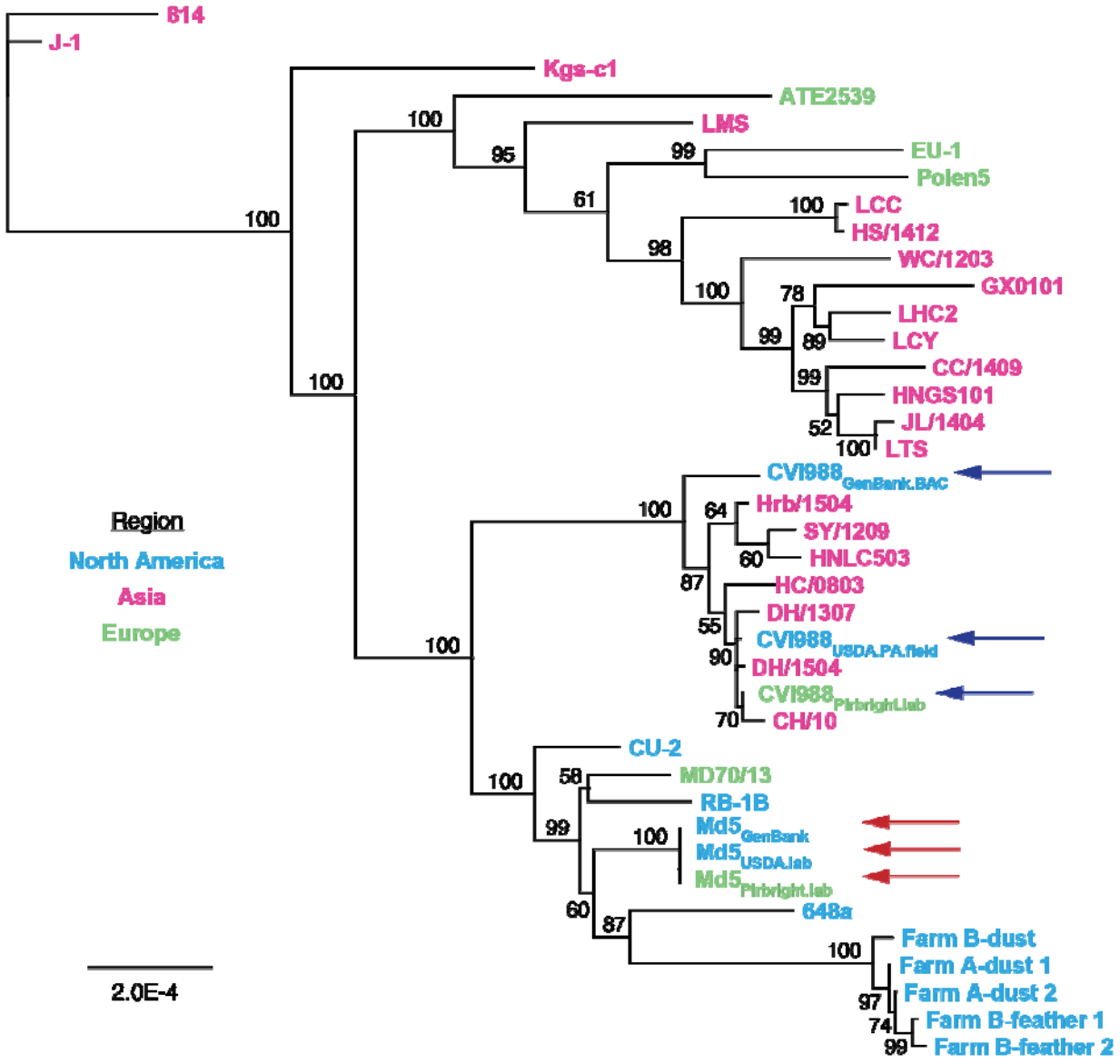
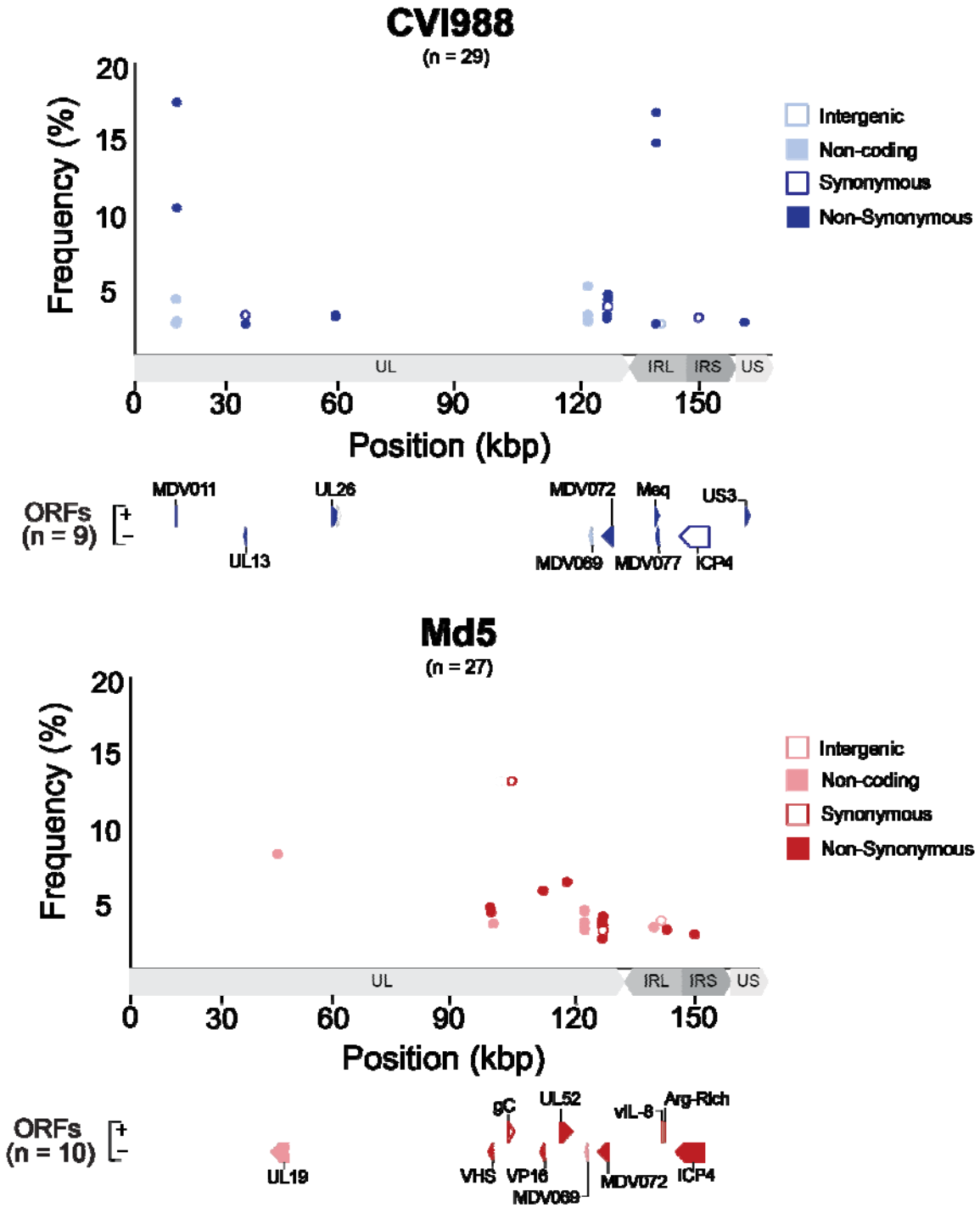


Figure 4: Minor variant distribution in Pirbright CVI988 and Md5 populations



Supplementary Tables

Supplementary Table 1: Variable positions in CVI988 consensus genomes

Genome	ORF	Alternative name	Position	This CVI988 genome	Other CVI988 genomes	Description
USDA.PA.field	MDV034	UL22 (gH)	52,857	T	C	Synonymous
GenBank.BAC	MDV062	UL49 (VP22)	111,859	C	A	Synonymous
GenBank.BAC	MDV062	UL49 (VP22)	111,866	C	G	Ser155 -> Cys155
GenBank.BAC	MDV062	UL49 (VP22)	111,876	G	C	Ala152 -> Pro152
USDA.PA.field	MDV074	-	127,352	C	-	Frameshift
USDA.PA.field	MDV074	-	127,353	G	-	Frameshift
USDA.PA.field	MDV074	-	127,354	G	-	Frameshift
USDA.PA.field	MDV074	-	127,355	C	-	Frameshift
USDA.PA.field	MDV074	-	127,356	T	-	Frameshift
Pirbright.lab	MDV075	-	127,855	T	C	Non-coding
GenBank.BAC	MDV075	-	130,654	C	-	Homopolymer extension, non-coding
GenBank.BAC	MDV075	-	130,655	C	-	Homopolymer extension, non-coding
Pirbright.lab	MDV075	-	128,849	-	C	Homopolymer contraction, non-coding
USDA.PA.field	MDV075	-	128,518	T	-	Homopolymer extension, non-coding
USDA.PA.field	MDV075	-	128,519	T	-	Homopolymer extension, non-coding
USDA.PA.field	MDV075	-	129,188	-	T	Homopolymer contraction, non-coding
GenBank.BAC	MDV075	-	131,742	A	T	Non-coding
GenBank.BAC	MDV075	-	131,766	C	-	Homopolymer extension, non-coding
GenBank.BAC	MDV075	-	131,978	A	-	Homopolymer extension, non-coding
GenBank.BAC	MDV075	-	131,979	A	-	Homopolymer extension, non-coding
Pirbright.lab	MDV075	-	130,169	-	A	Homopolymer contraction, non-coding
GenBank.BAC	-	-	132,535	T	-	Homopolymer extension
GenBank.BAC	-	-	132,536	T	-	Homopolymer extension
GenBank.BAC	-	-	134,710	A	-	Homopolymer extension
Pirbright.lab	-	-	133,116	-	C	Homopolymer contraction
USDA.PA.field	-	-	133,532	G	-	Homopolymer extension
Pirbright.lab	-	-	133,949	-	G	Homopolymer contraction
Pirbright.lab	MDV076	Meq	134,771	-	G…G	Meq isoform lacking 177 bp insertion
GenBank.BAC	MDV076	Meq	137,580	G	-	Homopolymer extension, non-coding
USDA.PA.field	MDV076	Meq	135,598	-	G	Homopolymer contraction, non-coding
Pirbright.lab	MDV076	Meq	135,686	T	-	Homopolymer extension, non-coding
GenBank.BAC	MDV076	Meq	137,670	-	TT	Homopolymer contraction, non-coding
GenBank.BAC	-	-	138,647	T	-	Homopolymer extension
GenBank.BAC	MDV078	vIL-8	140,205	A	-	Homopolymer extension, frameshift
USDA.PA.field	MDV080	-	138,849	-	T	Homopolymer contraction, frameshift
GenBank.BAC	-	-	141,642	A	G	Intergenic
GenBank.BAC	MDV085	-	152,337	C	-	Homopolymer extension, frameshift
GenBank.BAC	MDV085	-	152,351	C	A	Synonymous
GenBank.BAC	MDV086	-	153,340	A	-	Homopolymer extension, frameshift
Pirbright.lab	-	-	151,781	-	T	Homopolymer contraction
GenBank.BAC	-	-	154,397	A	-	Homopolymer extension
USDA.PA.field	-	-	152,409	-	G	Homopolymer contraction
GenBank.BAC	MDV091	US2	159,254	G	A	Val51 -> Ala51

GenBank.BAC	MDV092	US3	159,527	C	G	Ser4 -> Thr4
GenBank.BAC	MDV092	US3	159,552	T	C	Synonymous
GenBank.BAC	MDV092	US3	160,197	T	C	Synonymous
GenBank.BAC	MDV093	SORF4	160,899	C	T	Leu22->Ser22
GenBank.BAC	MDV093	SORF4	161,025	G	A	Tyr64 -> Cys64
GenBank.BAC	MDV093	SORF4	161,269	G	A	Synonymous
GenBank.BAC	MDV094	US6 (gD)	161,518	T	C	Synonymous
GenBank.BAC	MDV094	US6 (gD)	161,531	C	T	Ile29 -> Thr29
GenBank.BAC	MDV094	US6 (gD)	161,709	T	C	Synonymous
GenBank.BAC	MDV094	US6 (gD)	161,799	A	C	Synonymous
GenBank.BAC	MDV094	US6 (gD)	162,018	T	A	Synonymous
GenBank.BAC	MDV094	US6 (gD)	162,065	G	A	Glu207 -> Gly207
GenBank.BAC	MDV094	US6 (gD)	162,163	C	G	Gly240 -> Arg240
GenBank.BAC	MDV094	US6 (gD)	162,183	C	T	Synonymous
GenBank.BAC	MDV094	US6 (gD)	162,218	A	T	Ile258 -> Asn258
GenBank.BAC	-	-	166,119	C	A	Intergenic

Supplementary Table 2: Variable positions in Md5 consensus genomes

Genome	ORF	Position	Consensus	Call	Description
USDA.lab	MDV075	128,135	C	-	Homopolymer contraction, non-coding
GenBank	MDV075	129,507	-	A	Homopolymer extension, non-coding
GenBank	MDV075	129,508	-	A	Homopolymer extension, non-coding
GenBank	MDV075	129,588	-	T	Homopolymer extension, non-coding
GenBank	MDV075	130,253	-	T	Homopolymer extension, non-coding
GenBank	-	134,602	-	G	Homopolymer extension
USDA.lab	-	133,238	G	-	Homopolymer contraction
Pirbright.lab	-	152,552	C	-	Homopolymer contraction
Pirbright.lab	-	152,712	TG	-	Homopolymer contraction
Pirbright.lab	-	152,724	T	-	Homopolymer contraction

681 **Supplementary Table 3: Minor variants in CVI988$_{\text{Pirbright.lab}}$**

Position in the genome	Major allele	Minor allele	Minor allele frequency	Major allele # forward reads	Major allele # reverse reads	Minor allele # forward reads	Minor allele # reverse reads	ORF	Type
16732	T	C	4.65%	5651	2646	275	130	MDV011	Upstream non-coding
16735	G	T	2.22%	5817	2636	126	66	MDV011	Upstream non-coding
16770	A	G	3.21%	5990	6705	199	222	MDV011	Upstream non-coding
16809	A	G	17.30%	4691	5465	919	1206	MDV011	Non-synonymous (Thr2 -> Ala2)
16918	A	G	10.06%	2353	2384	266	264	MDV011	Non-synonymous (Glu38->Gly38)
36461	A	T	2.65%	2845	3416	63	108	MDV025	Synonymous
36463	T	G	2.09%	2853	3418	46	88	MDV025	Non-synonymous (Thr265->Pro265)
59022	G	T	2.55%	4836	3741	72	153	MDV038	Non-synonymous (Ala239->Ser239)
59040	A	C	2.70%	4512	2768	61	141	MDV038	Non-synonymous (Thr245->Pro245)
122353	G	A	2.11%	6326	5563	138	119	MDV069	Upstream non-coding
122362	G	A	2.26%	6323	5494	151	123	MDV069	Upstream non-coding
122405	A	G	2.34%	5802	5540	145	128	MDV069	Upstream non-coding
122419	T	C	2.66%	3553	5549	127	122	MDV069	Upstream non-coding
122421	T	G	2.66%	3552	5519	122	126	MDV069	Upstream non-coding
122848	A	G	2.43%	4877	5946	116	154	MDV072	Non-synonymous (Leu875->Pro875)
122857	A	G	2.66%	4802	5647	123	163	MDV072	Non-synonymous (Val872->Ala872)
122900	G	A	3.76%	3676	4473	141	178	MDV072	Non-synonymous (His858->Tyr858)
122914	C	T	3.25%	3332	4093	81	169	MDV072	Non-synonymous (Cyst853->Tyr853)
122916	G	T	3.27%	3270	4023	80	167	MDV072	Synonymous
122977	T	G	4.08%	2886	2890	84	162	MDV072	Non-synonymous (His832->Pro832)
123012	C	T	3.28%	2684	2155	51	113	MDV072	Non-synonymous (Trp820->STOP)
134751	A	G	14.59%	25181	19947	5447	2316	MDV076	Non-synonymous (Thr180->Ala180)
134759	A	G	16.71%	23815	19819	6391	2431	MDV077	Non-synonymous (Phe111->Leu111)
134784	C	-	58.08%	29714	19086	14302	14301	MDV076	Deletion
134785	C	-	57.02%	30172	19826	14287	14286	MDV076	Deletion
134786	A	-	57.12%	29125	19029	14288	14287	MDV076	Deletion
137914	T	C	2.20%	3776	3761	86	84	-	Intergenic
144482	T	C	2.60%	11669	13033	441	219	MDV084	Synonymous
157396	A	G	2.16%	6886	5749	157	123	MDV092	Non-synonymous (Glu54->Gly54)

682 **Supplementary Table 4: Minor variants in Md5$_{Pirbright.lab}$**

Position in the genome	Major allele	Minor allele	Minor allele frequency	Major allele # forward reads	Major allele # reverse reads	Minor allele # forward reads	Minor allele # reverse reads	ORF	Type
49848	A	G	7.89%	3400	3374	278	303	MDV031	Non-coding
97742	T	G	4.15%	2275	1949	107	76	MDV054	Non-synonymous (His136->Pro136)
97829	G	T	3.82%	2575	1105	106	40	MDV054	Non-synonymous (Thr107->Asn107)
98543	T	C	3.09%	1091	3673	33	119	MDV054	Non-coding
102545	A	G	12.92%	3903	2728	590	397	MDV057	Synonymous
110386	C	T	5.32%	4509	2255	259	121	MDV061	Non-synonymous (Asp35->Asn35)
114480	G	A	5.96%	2422	4690	229	222	MDV066	Non-synonymous (Val192->Ile192)
122515	G	A	2.61%	4183	3701	53	158	MDV069	Non-coding
122524	G	A	2.80%	4122	3688	65	160	MDV069	Non-coding
122567	A	G	3.17%	3769	3694	68	177	MDV069	Non-coding
122581	T	C	3.90%	2128	3659	63	172	MDV069	Non-coding
122583	T	G	4.03%	2122	3624	63	179	MDV069	Non-coding
122645	G	T	3.05%	2982	3659	51	158	MDV069	Non-coding
122977	G	A	2.01%	3501	4391	92	70	MDV072	Non-synonymous (Ala886->Val886)
123010	A	G	2.97%	3503	4076	142	90	MDV072	Non-synonymous (Leu875->Pro875)
123019	A	G	3.29%	3495	3995	158	97	MDV072	Non-synonymous (Val872->Ala872)
123062	G	A	3.57%	2986	3374	151	85	MDV072	Non-synonymous (His858->Tyr858)
123076	C	T	2.55%	2852	3217	78	81	MDV072	Non-synonymous (Cyst853->Tyr853)
123078	G	T	2.57%	2803	3185	76	82	MDV072	Synonymous
123139	T	G	3.58%	2753	2417	110	82	MDV072	Non-synonymous (His832->Pro832)
123174	C	T	2.87%	2726	2034	90	51	MDV072	Non-synonymous (Trp820->STOP)
132447	C	T	2.68%	6309	4595	161	139	-	Intergenic
137705	G	A	3.12%	5261	3764	168	123	MDV078	Synonymous
138785	G	C	2.55%	158	301	4	8	MDV079	Non-synonymous (Ala107->Pro107)
145343	T	G	2.25%	6893	3297	137	98	MDV084	Non-synonymous (Glu1725->Ala1725)
147827	G	A	21.44%	3490	2598	963	699	MDV084	Non-synonymous (Ala897->Val897)
150329	T	C	31.19%	7233	6650	3294	3002	MDV084	Non-synonymous (Gln63->Arg63)

Supplementary Table 5: Published reference genomes for additional 33 strains used to construct NJ tree

Strain	Accession	Publication Date	Length	PMID	Region
Kgs-c1	LC589272	15-Oct-20	174,999	33228722	Asia
J-1	KU744555	18-Feb-17	176,118	27112385	Asia
814	JF742597	23-Feb-12	172,541	21984218	Asia
MD70/13	MF431495	12-Dec-17	177,844	29151863	Europe
648a	JQ806361	4-Apr-13	176,080	22923089	North America
RB-1B	EF523390	29-Oct-07	178,246	17721813	North America
CU-2	EU499381	5-Aug-08	176,922	18351449	North America
HNGS101	MG432697	6-Mar-18	175,888	29407378	Asia
JL/1404	KU744559	18-Feb-17	176,083	27112385	Asia
LTS	KU744557	18-Feb-17	176,023	27112385	Asia
CC/1409	KU744560	18-Feb-17	175,561	27112385	Asia
GX0101	JX844666	3-Dec-12	178,101	23166235	Asia
LHC2	MW247189	19-Apr-22	177,327	35266322	Asia
WC/1203	KU744558	18-Feb-17	176,057	27112385	Asia
LCY	KX290013	12-Apr-17	175,319	28194622	Asia
LCC	KU744556	18-Feb-17	175,525	27112385	Asia
HS/1412	KU744561	18-Feb-17	175,532	27112385	Asia
Polen5	MF431496	12-Dec-17	177,821	29151863	Europe
EU-1	MF431494	12-Dec-17	177,828	29151863	Europe
ATE2539	MF431493	12-Dec-17	177,868	29151863	Europe
LMS	JQ314003	1-Aug-12	177,526	22476905	Asia
HNLC503	MG518371	29-Apr-19	178,195	31698116	Asia
HC/0803	MW531728	10-Jul-22	174,808	35266322	Asia
SY/1209	MW247208	19-Apr-22	177,240	35266322	Asia
Hrb/1504	MW247204	19-Apr-22	177,839	35266322	Asia
DH/1307	MW247201	19-Apr-22	177,235	35266322	Asia
DH/1504	MW247202	19-Apr-22	177,991	35266322	Asia
CH/10	MW247200	19-Apr-22	177,401	35266322	Asia
Farm A-dust 2	KU173115	04-Jul-16	178,049	27747299	North America
Farm A-dust 1	KU173116	04-Jul-16	177,967	27747299	North America
Farm B-feather 1	KU173117	04-Jul-16	178,327	27747299	North America
Farm B-feather 2	KU173118	04-Jul-16	178,540	27747299	North America
Farm B-dust	KU173119	04-Jul-16	178,169	27747299	North America