Targeted genomic enrichment and sequencing of CyHV-3 from carp tissues confirms low nucleotide diversity and mixed genotype infections

Saliha Hammoumi[1], Tatiana Vallaeys[2], Ayi Santika[3], Philippe Leleux[4], Ewa Borzym[5], Christophe Klopp[4], Jean-Christophe Avarre[1]

[1]Institut des Sciences de l'Evolution de Montpellier, IRD, Montpellier, France
[2]Université de Montpellier, Montpellier, France
[3]Main Center for Freshwater Aquaculture Development, Sukabumi, Indonesia
[4]Plate-forme Genotoul Bioinfo, UR875 Biométrie et Intelligence Artificielle, INRA, Castanet-Tolosan, France
[5]National Veterinary Research Institute, Pulawy, Poland

Correspondence:
Dr Jean-Christophe Avarre
jean-christophe.avarre@ird.fr

Running title: Targeted Genomic enrichment of CyHV-3

Abstract

Koi herpesvirus disease (KHVD) is an emerging disease that causes mass mortality in koi and common carp, *Cyprinus carpio* L. Its causative agent is Cyprinid herpesvirus 3 (CyHV-3), also known as koi herpesvirus (KHV). Although data on the pathogenesis of this deadly virus is relatively abundant in the literature, still little is known about its genomic diversity and about the molecular mechanisms that lead to such a high virulence. In this context, we developed a new strategy for sequencing full-length CyHV-3 genomes directly from infected fish tissues. Total genomic DNA extracted from carp gill tissue was specifically enriched with CyHV-3 sequences through hybridization to a set of nearly 2 million overlapping probes designed to cover the entire genome length, using KHV-J sequence (GenBank accession number AP008984) as reference. Applied to 7 CyHV-3 specimens from Poland and Indonesia, this targeted genomic enrichment enabled recovery of the full genomes with >99.9% reference coverage. The enrichment rate was directly correlated to the estimated number of viral copies contained in the DNA extracts used for library preparation, which varied between ~5000 to ~$2x10^7$. The average sequencing depth was >200 for all samples, thus allowing the search for variants with high confidence. Sequence analyses highlighted a significant proportion of intra-specimen sequence heterogeneity, suggesting the presence of mixed infections in all investigated fish. They also showed that inter-specimen genetic diversity at the genome scale was very low (> 99.95% of sequence identity). By enabling full genome comparisons directly from infected fish tissues, this new method will be valuable to trace outbreaks rapidly and at a reasonable cost, and in turn to understand the transmission routes of CyHV-3.

Introduction

Koi herpesvirus disease (KHVD) is an emerging disease that causes mass mortalities in koi and common carp, *Cyprinus carpio* L. Since the first report in the late 1990s (Hedrick et al. 2000), the disease has spread to many countries worldwide (Pokorova et al. 2005). The causative agent, cyprinid herpesvirus 3 (CyHV-3), is a large double-stranded DNA virus belonging to the *Alloherpesviridae* family, genus *Cyprinivirus* (Davison 2010; King et al. 2011; Waltzek et al. 2005). Twelve alloherpesviruses have been described to date: two of them infect amphibians while the ten remaining infect various fish species, including eel, salmon, catfish, goldfish and sturgeon (Waltzek et al. 2009).

Although the pathogenesis of this virus is well understood and well described in the literature, little is known about the genomic diversity and evolution. Sequencing of three complete CyHV-3 genomes of isolates originating from the USA (U), Israel (I) and Japan (J) (Aoki et al. 2007) revealed that sequences were highly similar (>99.9%), which was consistent with a scenario of a unique virus that spread worldwide. Meanwhile, accurate genomic comparisons highlighted two main genetic lineages formed by the U/I and J strains. The existence of two lineages was further confirmed on a larger set of European and Asian isolates by using PCR-based molecular markers targeting various regions of the genome (Bigarré et al. 2009; Kurita et al. 2009). These two lineages were designated as European and Asian. Finally, the use of variable number of tandem repeats (VNTR) revealed a whole range of variants associated with both European and Asian lineages (Avarre et al. 2011). It was also demonstrated that genetically distinct strains of CyHV-3 may co-occur in a single enzootic area, and that carp were frequently infected with a mix of genotypes (Avarre et al. 2012; Sunarto et al. 2011). Although viruses evolve more rapidly than their hosts, the differentiation ability of CyHV-3 remains mostly unknown, and consequently, it is difficult to predict whether the presence of distinct genotypes within a single host results from distinct infection events, from the evolution of a unique strain, or from both. Thus, characterizing these genetic variants at the genome level is critical to better understand the ecological and evolutionary significance of such mixed infections, especially with regards to the virulence and the modes of propagation of the virus.

In this context, the objective of the present work was to develop a method enabling the sequencing of CyHV-3 full-length genomes directly from their host. Targeted genomic enrichment (TGE), or targeted sequence capture prior to sequencing, has been described for a large panel of applications (Mertes et al. 2011). In spite of its great potential in virology, the application of this approach has remained limited since its first description and successful application to human herpesviruses (Depledge et al. 2011). Using a set of probes designed to specifically capture CyHV-3 sequences, seven genomes were fully sequenced from carp gill tissue samples collected during several outbreaks. In order to reduce the cost associated with TGE, the influence of pre-capture multiplexing on the ability to recover full genomes and to discover mutations in CyHV-3 with high confidence was also evaluated.

Material and Methods

CyHV-3 specimens and DNA extraction
Seven CyHV-3 specimens were used in the present study. They consisted of gill tissues collected from infected carps and stored in 96% EtOH at -20°C. Six of them (J1_101110, CB4_181110, PP3_070411, I_10-3, I_09-2i3, J2_101110) were obtained from moribund cultured common

97   carps collected in Indonesia (Sukabumi district, West Java), during different outbreaks (Table 1),
98   whereas the seventh (PoB3) was obtained from a common carp collected during an outbreak in
99   Poland. Total genomic DNA was extracted from a small gill fragment with the DNAzol kit (Life
100   technologies) and the Wizard Genomic DNA purification kit (Promega) for the Indonesian and
101   Polish specimens, respectively, following the appropriate protocols provided by the
102   manufacturers. DNA concentration was measured by fluorometry (Qbit, Life Technologies), and
103   its integrity was verified by capillary electrophoresis (Bioanalyzer 2100, Agilent). The viral
104   DNA concentration was determined in each sample by real-time PCR, using the same method as
105   previously described (Avarre et al. 2011; Yuasa et al. 2005).
106
107   *CyHV-3 genome enrichment and sequencing*
108   About 2 µg of total DNA were used to establish the genomic libraries of each sample, using the
109   TruSeq DNA sample preparation kit (Illumina). DNA was sonicated using a Bioruptor device
110   (Diagenode). The average insert size used for sequencing was ~300 bp. After adapter ligation
111   using the Truseq kit, libraries were subjected to a viral-specific sequence capture using the
112   SeqCap EZ library SR technology (Roche-Nimblegen). A large set of 50-105mer CyHV-3
113   specific probes was designed (Roche-Nimblegen) to theoretically capture all DNA sharing more
114   than 90% identity with the published CyHV-3 genome of strain J (GenBank #AP008984). This
115   probe set (Roche Diagnostics design OID 41307, IRN 4000012980) is available upon demand.
116   A pre-capture ligation-mediated PCR (LM-PCR) consisting of 8 cycles of PCR was first
117   performed according to the manufacturer's instructions (Illumina). After purification using
118   Agencourt Ampure XP beads (Beckman Coulter), the amplified products were hybridized with
119   the CyHV-3 probes at 47°C for 64 hours. In 4-plex assays, 250 ng of four amplified products
120   were pooled prior to hybridization. At the end of incubation, hybridized DNA was recovered
121   using Nimblegen capture beads. Then 18 cycles of a post-capture LM-PCR were carried out on
122   each captured DNA sample. DNA was again purified using Ampure beads, and its concentration
123   was measured using a spectrophotometer (Nanodrop 1000) while its integrity was verified using
124   a Bioanalyzer 2100 (Agilent).
125   In view of reducing the experimental costs, we compared the efficiency of a pre-capture
126   multiplexing (4-plex) design to that of a simplex one, using specimen PoB3 as "control" because
127   this sample displayed the lowest viral load (Table 1). Each library had initially been ligated with
128   a unique Illumina adaptor, and two 4-plex assays were designed according to the estimated viral
129   loads: the first included PoB3 and 3 Indonesian specimens with high viral loads ($>10^6$ genome
130   copies) whereas the second comprised PoB3 and the 3 remaining Indonesian specimens that
131   displayed lower loads ($\leq 1.5 \times 10^5$ genome copies). In parallel, PoB3 was also subjected to a
132   double-capture protocol, which consisted of two overnight hybridizations at 47°C followed by a
133   non-specific DNA elution using capture beads. All sample combinations were sequenced on a
134   single lane of a HiSeq2000 platform (Illumina) in a paired-end (100 bases) format, at
135   Montpellier Genomix (Montpellier, France).
136
137   *Genome assembly/mapping and detection of variations*
138   Raw sequences (fastq files) were stored in the public Sequence Read Archive (SRA) repository
139   under the SRP study accession SRP05976 as well as in the NG6 repository (Mariette et al. 2012)
140   of GenoToul (Toulouse, France) for analysis purposes. A first quality control consisted in the
141   visual   inspection   of   the   different   graphs   produced   by   fastqc
142   (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/), as well as in the alignment of the

143 reads against classical contaminants such as *Escherichia coli*, yeast and phages. All the raw reads
144 which passed this quality control were aligned to the CyHV-3 J strain genome (accession number
145 AP008984) using BWA MEM (version bwa-0.7.5a) (Li & Durbin 2009). Low quality
146 alignments, *i.e.* reads aligning more than once or with poor sequencing quality inducing
147 alignment quality lower than 30, were filtered out. The read depth was computed for the
148 reference J-strain genome using Samtools (version 1.1) (Li et al. 2009). Variant calling was
149 performed on the filtered alignments using the Genome Analyses Tool Kit (GATK, version
150 3.3.0) (DePristo et al. 2011), which improves the variant calling process through read
151 realignment and base quality recalibration. PCR duplicates were filtered out using picard tools
152 MarkDuplicates (version 1.88). Variations were thus obtained for single nucleotide
153 polymorphisms (SNPs), insertions and deletions (indels) in VCF format files, using a Q20 filter.
154 Both mapping and variant calling results were manually checked using the Integrative Genomics
155 Viewer (IGV, version 2.3) (Thorvaldsdottir et al. 2013). Finally, regions showing high coverage
156 were checked for contamination using blastn with standard settings (Altschul et al. 1997) against
157 the NCBI nucleotide collection (nr/nt) database. Figures representing the sequencing depth and
158 the number of small nucleotide variants were created with the R packages ggplot2 and limma,
159 respectively.
160 A *de novo* assembly of the reads was concomitantly performed for each sample using ABySS,
161 and Cap3 programs. Briefly, fastq files were trimmed with Fastq trimmer (version 1.0.0)
162 (Blankenberg et al. 2010) using the default parameters and the resulting reads were assembled in
163 paired-end by ABySS v1.5.0 (Simpson et al. 2009). The obtained contigs were subjected to a
164 new assembly using Cap3 (Huang & Madan 1999), and resulting assemblies were mapped on a
165 CyHV-3 reference genome (strain U, accession number DQ657948) using Bowtie 2 (version
166 2.2.1) program (Langmead et al. 2009). Consensus sequences were generated from these
167 assemblies, in which indels and SNPs were validated when more than 50% of the reads harbored
168 the variation.
169
170 *Genome comparisons*
171 Read alignments were also subjected to a genetic distance analysis. Variant calling files of
172 individual specimens were compared through an identity-by-state calculation using PLINK, with
173 the distance option (Chang et al. 2015). Reads were artificially created for the 3 reference
174 genomes U, I and J in order to include them in the analyses. For this purpose, 100-base reads
175 were artificially produced every 5 bases along the reference genome with a simple python script.
176 These reads were then randomly shuffled to form a set of reads with an even 20x depth, using
177 samtools faidx (version 0.1.19). Then a hierarchical cluster analysis was computed with a
178 complete-linkage clustering method, using the standard parameters, and dendrograms were
179 plotted with the hclust R package. A visual representation of the distribution of nucleotide
180 variants (SNPs + indels) was also created with Circos software (Krzywinski et al. 2009), using
181 the genome of CyHV-3 J strain as reference. For clarity, only the genomes of PoB3_1,
182 J1_101110 and I09_2i3 were represented.
183 The seven consensus sequences were aligned together with the 3 already sequenced genomes (J,
184 U and I strains) using the online Multiple alignment program for amino acid or nucleotide
185 sequences (Mafft program version 7, http://mafft.cbrc.jp/alignment/server/). The resulting
186 alignments were visualized and manually corrected with GeneDoc version 2.7 (Nicholas et al.
187 1997). Then the evolutionary history was inferred by the Neighbor-Joining method (Saitou &
188 Nei 1987), and evolutionary distances were computed with MEGA6 (Tamura et al. 2013), using

189 the Maximum Composite Likelihood method (Tamura et al. 2004). All positions containing gaps
190 and missing data were eliminated.
191
192 *Nucleotide sequence accession numbers*
193 The raw sequences (fastq files) of the whole project are available under the SRP study accession
194 SRP059764 (http://trace.ncbi.nlm.nih.gov/Traces/sra/?study=SRP059764). It is organized in 7
195 biosamples and 10 experiments, as shown in Table 1. Sequences corresponding to the 7
196 consensus genomes have also been submitted to GenBank and can be accessed under the
197 following accession numbers: KX544842, KX544843, KX544844, KX544845, KX544846,
198 KX544847 and KX544848 (for PoB3, J2_101110, CB4_181110, J1_101110, I_09_2i3, I_10_3
199 and PP3_070411, respectively).
200
201 **Results**
202
203 *Viral genomic DNA enrichment: read mapping, sequencing depth and genome coverage*
204 A total of 316,007,054 reads of 100 bp were obtained for the ten libraries, resulting in about 31.6
205 Gb of sequence data. The number of reads per library ranged from 9,163,662 (PoB3_3) to
206 56,046,694 (I_09-2i3) (Table 2). The rate of enrichment, reflected by the proportion of mapped
207 reads, apparently correlated with the initial viral load, as it tended to increase concomitantly. For
208 specimens with viral loads $> 10^6$ copies, the proportion of mapped reads (after duplicate
209 removal) was comprised between 43.99% (PP3_070411) and 50.70% (CB4_181110).
210 Conversely, for PoB3, it varied between 3.22% and 7.22% of the total number of reads,
211 depending on the assay format. Removal of duplicate sequences generally led to a significant
212 reduction in the proportion of reads that could be mapped on the CyHV-3 reference genome
213 (strain J). Sequence depth along this reference genome was variable both across and within
214 specimens. It was > 2000x on average for all specimens with an estimated number of viral copies
215 $> 10^5$, and > 200x for all other specimens (Table 2 and Supplementary figure 1). For most
216 samples, there was an over representation of reads at positions 226,373 - 226,831 and 228,015 -
217 228,163 (according to KHV-J coordinates, GenBank # AP008984). A Blast search revealed that
218 these regions shared 97% and 94% identity with sequences of *Cyprinus carpio* genome
219 (GenBank # LN591727 and LN591823, respectively). However, this over representation was
220 attenuated in samples with high initial viral loads (J1_101110, CB4_181110 and PP3_070411).
221 A higher coverage (x30) was also noticed for J2_101110 specimen in the region corresponding
222 to positions 45,500 - 47,700; yet this portion of CyHV-3 genome did not return any significant
223 blast hit. However, since the genome of *Cyprinus carpio* is not yet complete, it cannot be ruled
224 out that these sequences belong to the carp genome. In spite of these slight differences, genomes
225 could be recovered in their totality (or near-totality) from all samples, since coverage along the
226 reference sequence was above 99.9%, with the exception of PoB3_2 for which coverage was
227 slightly lower (99.05%).
228
229 *Comparison of PoB3 sequences according to the assay format*
230 Comparison of PoB3 in the different situations suggests that pre-capture multiplexing had an
231 influence on the number of raw reads: the number of obtained sequences was highest when PoB3
232 was enriched separately, lowest when mixed with specimens with high viral loads (PoB3_3) and
233 intermediate when mixed with specimens displaying lower viral loads (PoB3_4). The proportion
234 of mapped reads (before duplicate removal) followed the same trend, with the noticeable

235   exception of PoB3_2 (89.40%) (Table 2 and Supplementary Figure 2). However, this high
236   proportion of mapped reads obtained with the double-capture protocol (PoB3_2) dropped to
237   4.59% after sequence duplicate removal, suggesting that the majority of these sequences
238   corresponded to duplicates. Even though the mean depth values of the 4 PoB3 samples were
239   comparable, their distribution along the genome, however, displayed marked differences:
240   whereas sequence representation was the steadiest along the genome for PoB3_1, it was uneven
241   for PoB3_2, and in a lesser extent for PoB3_3 and PoB3_4 (figure 1).
242   When compared to KHV-J reference strain, the number of variations (SNP + indel) observed for
243   the 4 PoB3 samples varied from 423 (PoB3_1) to 457 (PoB3_3). Among these variations, 335
244   were shared by all 4 PoB3 samples (Figure 2). PoB3_3 displayed the highest number of 'unique'
245   variations (75), representing ~16% of the total variations recorded for this sample. Examination
246   of these 75 positions revealed that they all were 'heterozygous', *i.e.* they consisted in a mix of
247   two haplotypes, including that of the reference genome. Fifty five of these variations were shared
248   with at least one of the 3 other co-enriched samples (J1-101110, CB4-181110, PP3_070411), and
249   10 additional ones were located in the region that matches with the host genome (between
250   positions 226,373 and 228,163 according to KHV-J coordinates). A distance analysis including
251   the four PoB3 replicates (Supplementary Figure 3) indicated that the impact of these differences
252   on the general phylogeny was very limited.
253
254   *Inter- and intra-specimen variations*
255   Using KHV-J as reference, sequence identity ranged from 99.85% to 99.97%. The most
256   divergent sequence was that of PoB3. This corresponded to approximately 425 sites that varied
257   from the reference KHV-J sequence. The proportion of single nucleotide polymorphisms (SNP)
258   and insertions/deletions (indel) was nearly 50%-50% for all samples (Table 3). As can be
259   visualized on figure 4, variations were distributed across the entire genome, with no particular
260   hotspot of variability. In addition to the differences between specimens, sequence variability also
261   occurred within specimens. Indeed, the proportion of varying sites harboring more than one
262   allele (using the GATK criteria) was above 30% for all specimens (Table 3). If some of this
263   intra-specimen variability may probably be attributed to experimental biases (see discussion
264   below), it nonetheless suggests the presence of multiple genotypes within a single specimen.
265   Because of this intra-specimen sequence heterogeneity, distance analysis between specimens was
266   realized directly from vcf files, using the PLINK software. Results indicated that PoB3 belonged
267   to the European lineage, whereas all Indonesian specimens belonged to the Asian one (Figure 3).
268   Multiple sequence alignment realized on the seven generated consensus genomes and the three
269   reference CyHV-3 genomes led to the same result (Supplementary Figure 4).
270
271   **Discussion**
272
273   Concentration of viral particles through cell culture propagation (providing a suitable cell line
274   is available for the investigated virus) is not always successful and requires days to obtain
275   sufficient viral titers. Using targeted sequence capture prior to sequencing enabled here to
276   recover ten complete (or near-complete) 295-kbp genomes in one single run of sequencing,
277   directly from infected tissues, and with mean read depth varying from 219x to 9179x. For
278   comparison, Liu et al. used 24 genomic DNA libraries of Epstein Barr virus (EBV) to recover
279   one full genome from a tumor tissue with an average read depth of 17x (Liu et al. 2011). This
280   illustrates the potential of TGE to overcome the high discrepancy in the proportions of cellular

281 and viral genomic DNA in raw samples obtained from infected tissues. Sample multiplexing is
282 now routinely employed in next-generation sequencing, through the use of specific barcodes
283 (Parameswaran et al. 2007). However, there are a limited number of studies describing pre-
284 capture multiplexing, though it may improve both efficiency and cost-effectiveness of TGE
285 (Shearer et al. 2012). In the present case, a rough estimation indicated that 4x multiplexing
286 resulted in a drop of the per sample price from ~1600€ to ~450€ (excluding sequencing costs).
287   Although there are no replicates to allow for statistical confirmation, the results obtained on
288 PoB3 samples revealed some interesting differences. First, the double-capture protocol resulted
289 in a higher number of uncovered positions, suggesting that the repeated amplifications have
290 generated a bias in sequence representation. Though this aspect is not well documented in the
291 literature, it has been shown, however, that the time of barcode incorporation during the library
292 preparation and its distance from the priming site, as well as the PCR extension temperature,
293 were potential sources of bias (López-Barragán et al. 2011; Van Nieuwerburgh et al. 2011).
294 Second, while pre-capture multiplexing entailed a decrease in the number of reads, it did not
295 compromise the sequencing depth or the coverage along the genome. Third, there were slight
296 differences in the number of variations found between the four PoB3 samples and KHV-J
297 reference genome, with PoB3_3 displaying the highest number of 'unshared' variations, *i.e.*
298 variations that were not recorded in the other PoB3 samples. It is interesting to note that the 75
299 variations that were specific to PoB3_3 were all heterozygous, and that the majority of them
300 were shared by at least one of its co-enriched counterparts. Of these 75 variants, only 10 were
301 located in the region that shares a high sequence identity with the carp reference genome. This
302 evidence is likely indicative of low level cross-contamination with co-enriched samples.
303 Comparison with PoB3_4, for which the 13 specific variations were also heterozygous and
304 shared by at least one of the multiplexed samples, suggests that such a contamination would be
305 favored by the difference in the initial viral loads of the multiplexed samples. For this reason, it
306 seems advisable to multiplex samples with comparable viral loads.
307   It is generally acknowledged that errors during the mapping of cluster coordinates may result
308 in 1-2% of bias in assignment of reads to sample during demultiplexing (Krueger et al. 2011).
309 Considering the great difference in the sequencing depth between PoB3 samples and the other
310 specimens, such a "contamination" could explain these faint differences between PoB3 samples.
311 Moreover, an additional source of inaccuracy may also emanate from amplification and/or
312 sequencing errors. In order to evaluate the contribution of pre-capture multiplexing on these
313 observed differences, a specimen with a high viral load ($1.5x10^8$ viral copies) was passaged on
314 CCB (common carp brain) cells and sequenced: 1) after TGE in a 4-plex format with specimens
315 of comparable viral loads; 2) without enrichment. Comparison of the resulting sequences
316 revealed only 18 differences over the 295-kbp genomes between the enriched and non-enriched
317 samples (data concerning this specimen and additional ones will be published separately). This
318 result brings additional evidence that pre-capture multiplexing does not induce any significant
319 bias as long as the multiplexed specimens have comparable viral loads, and/or as long as the
320 initial viral amount is sufficient to generate elevated read depth (>1000x). In any case, the
321 differences between the four PoB3 samples represented nearly 0.004 to 0.025% of the genome,
322 *i.e.* 4-25 differences every 100 kb. This error rate appears acceptable, especially if we consider
323 the low number of viral copies that were initially contained in this specimen.
324   Considering the relatively low viral load of J2_101110 specimen, we can also expect that a
325 number of observed variations may constitute 'false callings'. This can be anticipated by i) the
326 total number of detected variations (171), which is slightly higher than in the other Indonesian

327   specimens (100-126); and ii) the higher proportion of heterozygous variations compared with the
328   other Indonesian specimens. Even though the proportion of heterozygous positions is certainly
329   over-estimated, especially in samples that were multiplexed prior to TGE, such a high rate of
330   intra-specimen variability confirms previous findings that carp are frequently infected with a mix
331   of CyHV-3 genotypes (Sunarto et al. 2011). The way such mixed infections arise in their host
332   still remains speculative, as well as their ecological and evolutionary advantage. Now, the
333   possibility to access full viral genomes from their host will undoubtedly contribute to a better
334   understanding of CyHV-3 mixed infections.
335   The genomic distance analysis indicated that PoB3 specimen belonged to the European
336   lineage, whereas the Indonesian samples fell in the Asian lineage. However, as was formerly
337   shown (Aoki et al. 2007; Li et al. 2015), the divergence remains very low at the genome scale.
338   The first KHVD outbreak reported in Indonesia dates back to 2002 (Sunarto et al. 2002). Since
339   then, the disease rapidly spread throughout the whole country, especially in the West Java region
340   (Sunarto et al. 2005). The use of 12 VNTR (variable number of tandem repeat) markers on the
341   six investigated CyHV-3 Indonesian specimens revealed an important diversity of genotypes,
342   which is not reflected at the genome scale (Avarre et al. 2012). This is not surprising, as tandem
343   repeats evolve faster than any other part of a genome. Whereas some tandem repeats do not carry
344   any adaptive value, others, however, allow functional diversification and may be involved in
345   pathogenic pathways (Mrázek et al. 2007; Treangen et al. 2009). Therefore, full genomic
346   comparisons will probably be more suitable to trace back outbreaks and to understand the
347   transmission routes of CyHV-3, as they will also enable the study of the variations of short
348   tandem repeats contained in the genomes of CyHV-3.
349
350   **Conclusion**
351
352   This TGE strategy enabled the successful recovery of full-length genomes from samples
353   containing as little as 5000 viral copies. Though pre-capture multiplexing of samples with low
354   viral loads is prone to generate false variants, it dramatically reduces the cost of an experiment.
355   From our results, we would advise to pre-capture multiplex only samples with an initial number
356   of viral copies above 10^5, and to multiplex samples with comparable loads whenever possible.
357   This TEG strategy will enable to analyze the genomic diversity of CyHV-3 and to study the
358   dynamics of *in vivo* infections, by accessing the possible heterogeneity in genome populations
359   through the analysis of the nucleotide composition at each position along the genome.
360

361   **Acknowledgement**
362   This is publication IRD-DIVA-ISEM 2016-162.
363

364 **References**
365
366 Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, and Lipman DJ. 1997.
367   Gapped BLAST and PSI-BLAST: a new generation of protein database search programs.
368   *Nucleic Acids Research* 25:3389-3402.
369 Aoki T, Hirono I, Kurokawa K, Fukuda H, Nahary R, Eldar A, Davison AJ, Waltzek TB,
370   Bercovier H, and Hedrick RP. 2007. Genome sequences of three koi herpesvirus isolates
371   representing the expanding distribution of an emerging disease threatening koi and
372   common carp worldwide. *Journal of Virology* 81:5058-5065.
373 Avarre JC, Madeira JP, Santika A, Zainun Z, Baud M, Cabon J, Caruso D, Castric J, Bigarre L,
374   Engelsma M, and Maskur M. 2011. Investigation of Cyprinid herpesvirus-3 genetic
375   diversity by a multi-locus variable number of tandem repeats analysis. *J Virol Methods*
376   173:320-327.
377 Avarre JC, Santika A, Bentenni A, Zainun Z, Madeira JP, Maskur M, Bigarré L, and Caruso D.
378   2012. Spatio-temporal analysis of Cyprinid herpesvirus 3 genetic diversity at a local
379   scale. *Journal of Fish Diseases* 35:767-774.
380 Bigarré L, Baud M, Cabon J, Antychowicz J, Bergmann SM, Engelsma M, Pozet F, Reichert M,
381   and Castric J. 2009. Differentiation between Cyprinid herpesvirus type-3 lineages using
382   duplex PCR. *Journal of Virological Methods* 158:51-57.
383 Blankenberg D, Gordon A, Von Kuster G, Coraor N, Taylor J, and Nekrutenko A. 2010.
384   Manipulation of FASTQ data with Galaxy. *Bioinformatics* 26:1783-1785.
385 Chang C, Chow C, Tellier L, Vattikuti S, Purcell S, and Lee J. 2015. Second-generation PLINK:
386   rising to the challenge of larger and richer datasets. *GigaScience* 4:7.
387 Davison AJ. 2010. Herpesvirus systematics. *Veterinary Microbiology* 143:52-69.
388 Depledge DP, Palser AL, Watson SJ, Lai IY-C, Gray ER, Grant P, Kanda RK, Leproust E,
389   Kellam P, and Breuer J. 2011. Specific Capture and Whole-Genome Sequencing of
390   Viruses from Clinical Samples. *PLoS ONE* 6:e27805.
391 DePristo MA, Banks E, Poplin R, Garimella KV, Maguire JR, Hartl C, Philippakis AA, del
392   Angel G, Rivas MA, Hanna M, McKenna A, Fennell TJ, Kernytsky AM, Sivachenko
393   AY, Cibulskis K, Gabriel SB, Altshuler D, and Daly MJ. 2011. A framework for
394   variation discovery and genotyping using next-generation DNA sequencing data. *Nat*
395   *Genet* 43:491-498.
396 Hedrick RP, Gilad O, Yun S, Spangenberg JV, Marty GD, Nordhausen RW, Kebus MJ,
397   Bercovier H, and Eldar A. 2000. A herpesvirus associated with mass mortality of juvenile
398   and adult koi, a strain of common carp. *Journal of Aquatic Animal Health* 12:44-57.
399 Huang X, and Madan A. 1999. CAP3: A DNA Sequence Assembly Program. *Genome Research*
400   9:868-877.
401 King AMQ, Adams MJ, Carstens EB, and Lefkowitz EJ. 2011. Virus taxonomy: Ninth Report of
402   the International Committee on Taxonomy of Viruses. San Diego: Elsevier Academic
403   Press. p 1327.
404 Krueger F, Andrews SR, and Osborne CS. 2011. Large Scale Loss of Data in Low-Diversity
405   Illumina Sequencing Libraries Can Be Recovered by Deferred Cluster Calling. *PLoS*
406   *ONE* 6:e16607.
407 Krzywinski MI, Schein JE, Birol I, Connors J, Gascoyne R, Horsman D, Jones SJ, and Marra
408   MA. 2009. Circos: An information aesthetic for comparative genomics. *Genome*
409   *Research*.

410 Kurita J, Yuasa K, Ito T, Sano M, Hedrick RP, Engelsma MY, Haenen OLM, Sunarto A,
411   Kholidin EB, Chou HY, Tung MC, de la Pena L, Lio-Po G, Tu C, Way K, and Iida T.
412   2009. Molecular Epidemiology of Koi Herpesvirus. *Fish Pathology* 44:59-66.
413 Langmead B, Trapnell C, Pop M, and Salzberg S. 2009. Ultrafast and memory-efficient
414   alignment of short DNA sequences to the human genome. *Genome Biology* 10:R25.
415 Li H, and Durbin R. 2009. Fast and accurate short read alignment with Burrows-Wheeler
416   transform. *Bioinformatics* 25:1754-1760.
417 Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R,
418   and Genome Project Data Processing S. 2009. The Sequence Alignment/Map format and
419   SAMtools. *Bioinformatics* 25:2078-2079.
420 Li W, Lee X, Weng S, He J, and Dong C. 2015. Whole-genome sequence of a novel Chinese
421   cyprinid herpesvirus 3 isolate reveals the existence of a distinct European genotype in
422   East Asia. *Veterinary Microbiology* 175:185-194.
423 Liu P, Fang X, Feng Z, Guo YM, Peng RJ, Liu T, Huang Z, Feng Y, Sun X, Xiong Z, Guo X,
424   Pang SS, Wang B, Lv X, Feng FT, Li DJ, Chen LZ, Feng QS, Huang WL, Zeng MS, Bei
425   JX, Zhang Y, and Zeng YX. 2011. Direct sequencing and characterization of a clinical
426   isolate of Epstein-Barr virus from nasopharyngeal carcinoma tissue by using next-
427   generation sequencing technology. *J Virol* 85:11291-11299.
428 López-Barragán MJ, Quiñones M, Cui K, Lemieux J, Zhao K, and Su X-z. 2011. Effect of PCR
429   extension temperature on high-throughput sequencing. *Molecular and Biochemical
430   Parasitology* 176:64-67.
431 Mariette J, Escudie F, Allias N, Salin G, Noirot C, Thomas S, and Klopp C. 2012. NG6:
432   Integrated next generation sequencing storage and processing environment. *Bmc
433   Genomics* 13:462.
434 Mertes F, ElSharawy A, Sauer S, van Helvoort J, van der Zaag PJ, Franke A, Nilsson M,
435   Lehrach H, and Brookes AJ. 2011. Targeted enrichment of genomic DNA regions for
436   next-generation sequencing. *Briefings in Functional Genomics* 10:374-386.
437 Mrázek J, Guo X, and Shah A. 2007. Simple sequence repeats in prokaryotic genomes.
438   *Proceedings of the National Academy of Sciences* 104:8472-8477.
439 Nicholas KB, Nicholas HBJ, and Deerfield DW. 1997. GeneDoc: Analysis and visualization of
440   genetic variation. *EMBNEWNEWS* 4:14.
441 Parameswaran P, Jalili R, Tao L, Shokralla S, Gharizadeh B, Ronaghi M, and Fire AZ. 2007. A
442   pyrosequencing-tailored nucleotide barcode design unveils opportunities for large-scale
443   sample multiplexing. *Nucleic Acids Research* 35:e130.
444 Pokorova D, Vesely T, Piackova V, Reschova S, and Hulova J. 2005. Current knowledge on koi
445   herpesvirus (KHV): a review. *Veterinarni Medicina* 50:139-147.
446 Saitou N, and Nei M. 1987. The neighbor-joining method: a new method for reconstructing
447   phylogenetic trees. *Mol Biol Evol* 4:406-425.
448 Shearer AE, Hildebrand MS, Ravi H, Joshi S, Guiffre AC, Novak B, Happe S, Le Proust EM,
449   and Smith RJH. 2012. Pre-capture multiplexing improves efficiency and cost-
450   effectiveness of targeted genomic enrichment. *Bmc Genomics* 13:8.
451 Simpson JT, Wong K, Jackman SD, Schein JE, Jones SJM, and Birol İ. 2009. ABySS: A parallel
452   assembler for short read sequence data. *Genome Research* 19:1117-1123.
453 Sunarto A, McColl KA, Crane MS, Sumiati T, Hyatt AD, Barnes AC, and Walker PJ. 2011.
454   Isolation and characterization of koi herpesvirus (KHV) from Indonesia: identification of
455   a new genetic lineage. *Journal of Fish Diseases* 34:87-101.

456 Sunarto A, Rukyani A, and Itami T. 2005. Indonesian Experience on the Outbreak of Koi
457   Herpesvirus in Koi and Carp
458 (Cyprinus carpio). *Bull Fish Res Agen* Supplement No. 2,:15-21.
459 Sunarto A, Taukhid, Rukyani A, Koesharyani I, Supriyadi H, Huminto H, Agungpriyono DR,
460   Pasaribu FH, Widodo, Herdikiawan D, Rukmono D, and Prayitno SB. 2002. Field
461   investigations on a serious disease outbreak among koi and common carp (*Cyprinus
462   carpio*) in Indonesia. 5th Symposium on Diseases in Asian Aquaculture. Gold Cost,
463   Australia.
464 Tamura K, Nei M, and Kumar S. 2004. Prospects for inferring very large phylogenies by using
465   the neighbor-joining method. *Proceedings of the National Academy of Sciences of the
466   United States of America* 101:11030-11035.
467 Tamura K, Stecher G, Peterson D, Filipski A, and Kumar S. 2013. MEGA6: Molecular
468   Evolutionary Genetics Analysis version 6.0. *Mol Biol Evol* 30:2725-2729.
469 Thorvaldsdottir H, Robinson JT, and Mesirov JP. 2013. Integrative Genomics Viewer (IGV):
470   high-performance genomics data visualization and exploration. *Brief Bioinform* 14:178-
471   192.
472 Treangen TJ, Abraham AL, Touchon M, and Rocha EP. 2009. Genesis, effects and fates of
473   repeats in prokaryotic genomes. *FEMS Microbiol Rev* 33:539-571.
474 Van Nieuwerburgh F, Soetaert S, Podshivalova K, Ay-Lin Wang E, Schaffer L, Deforce D,
475   Salomon DR, Head SR, and Ordoukhanian P. 2011. Quantitative Bias in Illumina TruSeq
476   and a Novel Post Amplification Barcoding Strategy for Multiplexed DNA and Small
477   RNA Deep Sequencing. *PLoS ONE* 6:e26969.
478 Waltzek TB, Kelley GO, Alfaro ME, Kurobe T, Davison AJ, and Hedrick RP. 2009.
479   Phylogenetic relationships in the family Alloherpesviridae. *Diseases of Aquatic
480   Organisms* 84:179-194.
481 Waltzek TB, Kelley GO, Stone DM, Way K, Hanson L, Fukuda H, Hirono I, Aoki T, Davison
482   AJ, and Hedrick RP. 2005. Koi herpesvirus represents a third cyprinid herpesvirus
483   (CyHV-3) in the family Herpesviridae. *Journal of General Virology* 86:1659-1667.
484 Yuasa K, Sano M, Kurita J, Ito T, and Iida T. 2005. Improvement of a PCR method with the Sph
485   1-5 primer set for the detection of koi herpesvirus (KHV). *Fish Pathology* 40:37-39.

486   Table 1 Assay design for CyHV-3 sequence capture

Assay format	Specimen name	Estimated number of viral copies[a]	Country and year of collection	Reference	Experiment accession #
Simplex	PoB3_1	4.8E+3	Poland, 2013	This study	SRX1071741
Simplex, double capture	PoB3_2	4.8E+3	Poland, 2013	This study	SRX1071742
4-plex	PoB3_3	4.8E+3	Poland, 2013	This study	SRX1071743
	J1_101110	1.5E+7	Indonesia, 2010	Avarre et al, 2012	SRX1071745
	CB4_181110	1.9E+7	Indonesia, 2010	Avarre et al, 2012	SRX1071746
	PP3_070411	5.7E+6	Indonesia, 2011	Avarre et al, 2012	SRX1071747
4-plex	PoB3_4	4.8E+3	Poland, 2013	This study	SRX1071744
	I_10-3	1.5E+5	Indonesia, 2010	Avarre et al, 2011	SRX1071748
	I_09-2i3	1.5E+5	Indonesia, 2009	Avarre et al, 2011	SRX1071749
	J2_101110	8.8E+4	Indonesia, 2010	Avarre et al, 2012	SRX1071750

487   [a] Before capture of viral DNA
488
489

490   Table 2 Main features of enrichment and sequencing results

Specimen name	# raw reads	% mapped reads[a]	% mapped reads after duplicate removal[a]	Mean depth $(x)^{a, b}$	Coverage along the reference genome (%)[a]	Number of uncovered positions[a]	Estimated number of viral copies
PoB3_1	22 852 739	28.73	3.22	221	99.99	32	4.8E+3
PoB3_2	28 611 191	89.40	4.59	409	99.05	2600	4.8E+3
PoB3_3	9 163 662	16.67	7.22	219	99.92	218	4.8E+3
J1_101110	32 892 125	80.47	49.66	5931	99.94	168	1.5E+7
CB4_181110	49 788 380	88.81	50.70	9179	99.99	29	1.9E+7
PP3_070411	42 335 046	76.32	43.99	6741	99.96	115	5.7E+6
PoB3_4	16 182 800	33.78	4.90	246	100	0	4.8E+3
I_10_3	44 364 454	67.84	16.46	2601	100	0	1.5E+5
I_09_2i3	56 046 694	74.00	19.87	4015	100	0	1.5E+5
J2_101110	14 067 420	31.61	5.26	235	99.90	272	8.8E+4

491   [a] Only Q30 sequences were considered

492   [b] after duplicate removal

493 Table 3 Number of single nucleotide polymorphisms (SNPs) and insertions/deletions (indels) in
494 the different specimens compared to KHV-J

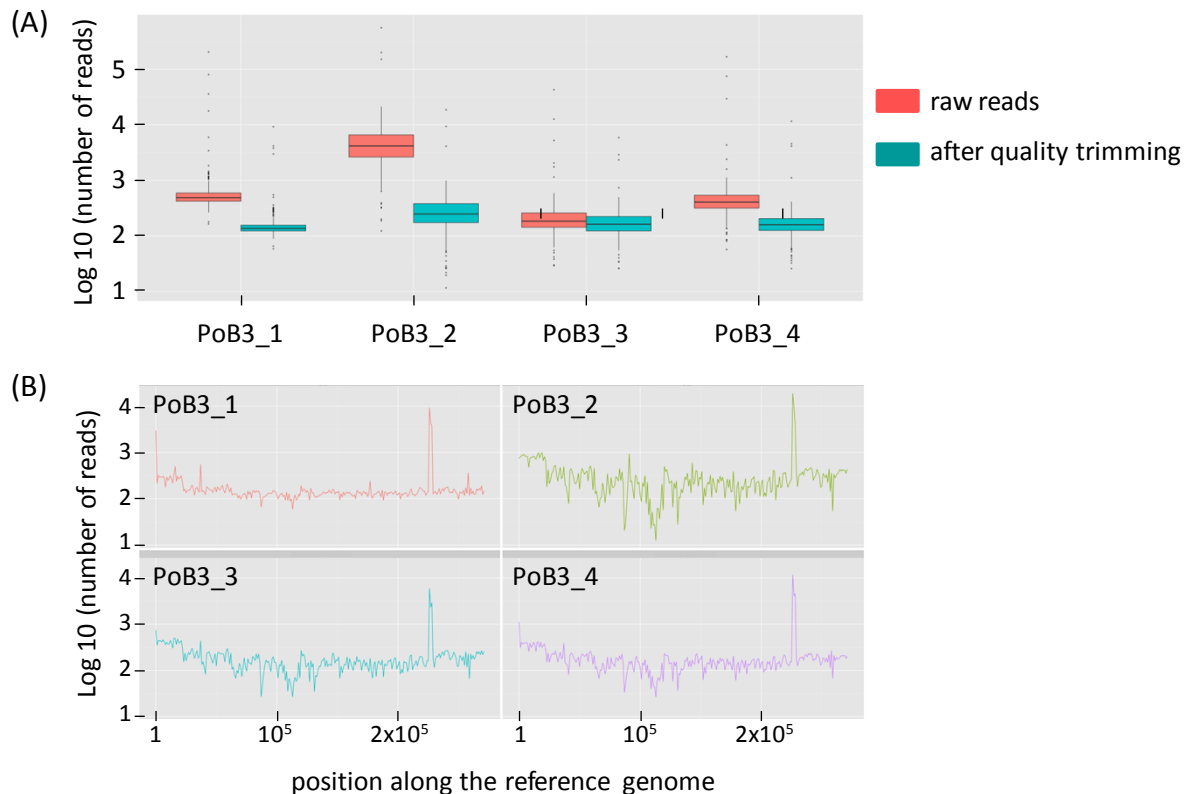
Specimen	SNP	Indel	% small nucleotide variants [a]	% intra-specimen heterogeneous sites[a, b]
PoB3_1	215	208	0.14	31.8
PoB3_2	216	208	0.14	33.6
PoB3_3	230	227	0.15	95.4
J1-101110	47	66	0.04	31.3
CB4-181110	43	57	0.03	31.3
PP3_070411	66	60	0.04	42.5
PoB3_4	213	211	0.14	93.0
I10-3	72	53	0.04	72.1
I_09-2i3	60	56	0.04	54.6
J2_101110	94	77	0.06	81.9

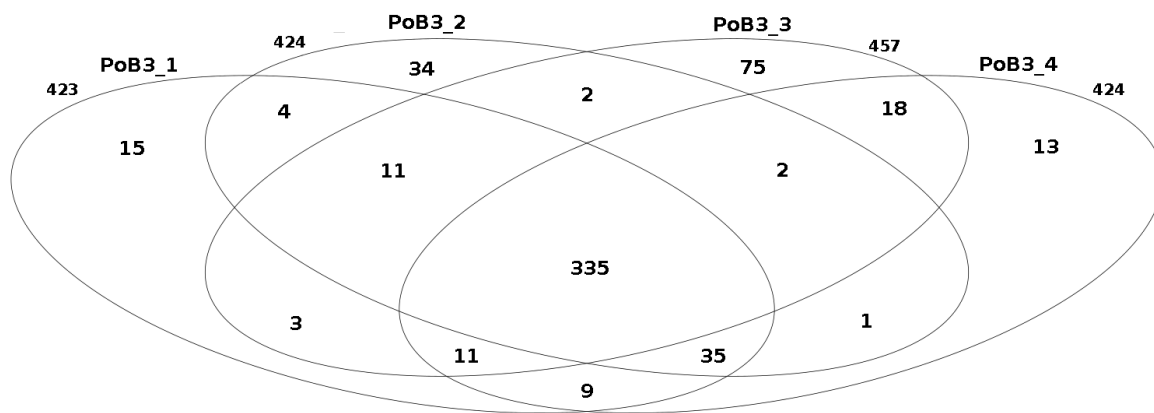
495 [a] Cut-off value set at 10% of the total number of reads at each specific position
496 [b] Excluding the terminal repeats
497

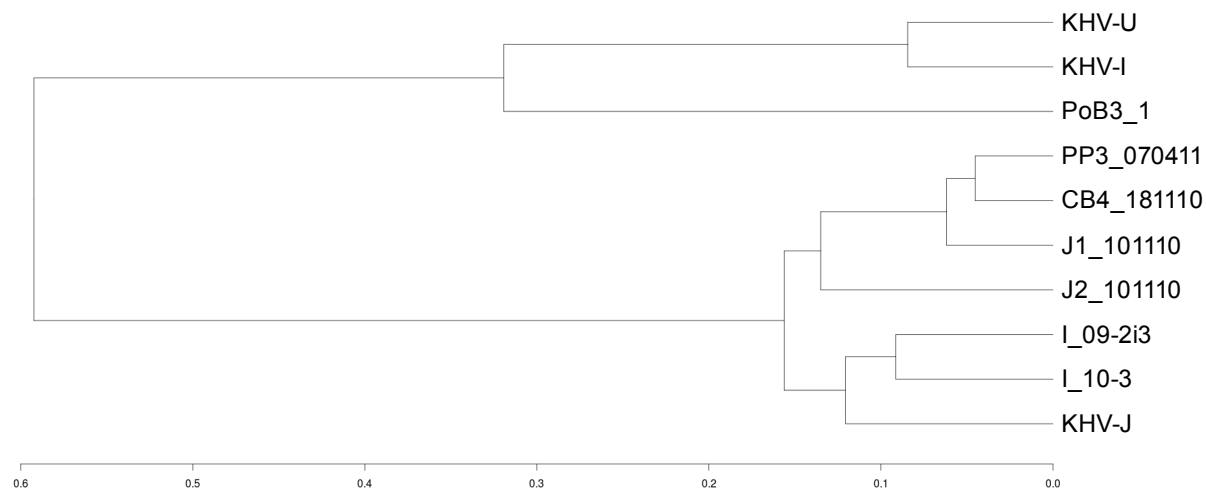
498 **Figure 1**: Sequencing depth for PoB3 samples. (A) Boxplot representing the depth distribution in
499 log10 of raw reads (red) and filtered reads (blue), *i.e.* after Q30 mapping quality trimming and
500 duplicate removal. (B) Per base sequencing depth along the KHV-J reference genome. The
501 figure was created with the R package ggplot2.
502

(A)

(B)

503
504

505 **Figure 2**: Comparison of the number of small nucleotide variants between PoB3 samples.
506 Intersections represent the variants that are shared between samples. This Venn diagram was
507 created with the R package limma.
508
509

510
511

512 **Figure 3**: Genetic distances between CyHV-3 specimens. A distance matrix between samples
513 was first computed using an Identity By State (IBS) distance. Specimens were then grouped into
514 clusters using the R package hclust and the corresponding dendrogram plotted. The scale
515 represents the maximum IBS-distance between two elements of linked clusters (branches in the
516 tree).
517

518
519

520 **Figure 4**: Distribution of nucleotide variations along CyHV-3 genomes. Numbers correspond to
521 the genomic coordinates of CyHV-3 J strain, and the internal curves depict the small nucleotide
522 variants (including SNPs and indels) found in the 3 genomes (PoB3_1, J1_101110 and I09_2i3)
523 compared to CyHV-3 J strain. Peak height is proportional to the number of variations in a given
524 region. A window size of 5kb was used.
525
526

527