JC Polyomavirus whole genome sequencing at the single molecule level reveals emerging neurotropic populations in Progressive Multifocal Leucoencephalopathy

Anne-Sophie L'honneur, Juliana Pipoli da Fonseca, Thomas Cokelaer, Flore Rozenberg

JC Polyomavirus whole genome sequencing at the single molecule level reveals emerging neurotropic populations in Progressive Multifocal Leucoencephalopathy

Anne-Sophie L'Honneur[1,2 *], Juliana Pipoli Da Fonseca[3], Thomas Cokelaer[3,4], Flore Rozenberg[1,2 *]

[1] *Université de Paris, INSERM Paris, France,* [2] *Assistance Publique-Hôpitaux de Paris, Hôpital Cochin, Service de Virologie, Paris, France,* [3] *Plate-forme Technologique Biomics – Centre de Ressources et Recherches Technologique (C2RT), Institut Pasteur, Paris, France,* [4] *Hub de Bioinformatique et de Biostatistique, Département Biologie Computationnelle, Institut Pasteur Paris, France*

*Address correspondence to flore.rozenberg@aphp.fr.

Correspondance address :

Flore Rozenberg

Service de Virologie

Hôpital Cochin, 27 rue du Faubourg Saint Jacques 75014 PARIS, FRANCE

Tél : 33 (0)1 48 51 33 19

Summary: Intra-host deep sequencing of JC Polyomavirus whole genome in Progressive Multifocal Leucoencephalopathy patients reveals compartment-specific and opposite patterns of genome variations in brain and urine. Neurotropic quasispecies stem from initial deletions in non-coding-control-region, evolve rapidly and acquire mutations in capsid.

Abstract

Background: JC polyomavirus (JCV) mostly causes asymptomatic persistent renal infections but may give rise in immunosuppressed patients to neurotropic variants which replicate in the brain causing progressive multifocal leukoencephalopathy (PML). Rearrangements in the JCV genome regulator non-coding control region (NCCR) and missense mutations in the viral capsid VP1 gene differentiate neurotropic variants from virus excreted in urine. **Methods**: To investigate intra-host emergence of JCV neurotropic populations in PML, we deep sequenced JCV whole genome recovered from cerebrospinal fluid (CSF) and urine samples from 32 HIV- and non HIV-infected PML patients at the single-molecule level. **Results**: JCV strains distributed among 6 out of 7 known genotypes. Common patterns of NCCR rearrangements included an initial deletion mostly located in a short 10-nucleotide sequence, followed by duplications/insertions. Multiple NCCR variants present in individual CSF samples shared at least one rearrangement suggesting they stemmed from a unique viral population. NCCR variants independently acquired single or double PML-specific adaptive VP1 mutations. NCCR variants recovered from urine and CSF displayed opposite deletion or duplication patterns in binding sites for transcription factors. **Discussion:** Long read deep sequencing shed light on emergence of neurotropic JCV populations in PML.

Key words

JC Polyomavirus; neurotropic; progressive multifocal leukoencephalopathy; long-read sequencing; NCCR rearrangement; VP1 mutation; deletion hotspot

Background

JC Polyomavirus (JCV) is a ubiquitous human virus causing asymptomatic persistent infections. In several immunosuppression conditions, JCV causes progressive multifocal leukoencephalopathy (PML), a rare but severe disease due to JCV replication in the brain. PML was described in hematologic malignancies**,** then strongly associated with human immunodeficiency virus (HIV) infection/acquired immune deficiency syndrome (AIDS). Novel causes of PML include immunomodulatory therapies, particularly the anti-integrin monoclonal antibody Natalizumab in multiple sclerosis patients [1]. The absence of antiviral drug directed to Polyomaviruses and the wide use of PML-causing drugs have generated new interest in the pathogenesis of PML.

PML results from lytic infection of oligodendrocytes by a neurotropic form of JCV. After primary infection, mostly acquired silently during childhood via presumed oropharyngeal exposure, the virus establishes a persistent infection in the kidney and is intermittently excreted in urine [2]. Other tissues have been reported as sites of persistence, such as lymph nodes [3], bone marrow [3], tonsils [4], and the brain [3,5–7]. High JCV replication levels in absence of immune control are thought to favour emergence of neuropathogenic JCV variants carrying structural genomic characteristics [8].

The JCV circular double-stranded DNA genome encodes early and late genes separated by the non-coding control region (NCCR), a 400 bp region including promoters and *cis*-regulating elements involved in viral replication and transcription. The NCCR sequence from urinary JCV strains displays a conserved "archetype" structure characterised by 6 sections named "*a*" to "*f*" [9,10]. Archetype strains detected in urine [11] and in environment [12] are probable sources of human contamination. In contrast, the first reported "prototype" CNS sequence from a PML patient, JCV "Mad-1", carried deleted *b* and *d* sections, and duplicated *a-c-e* sequences [13]. Subsequent analysis of PML CNS strains revealed that prototype

NCCR variability is unique to each patient, showing various deletions and duplications [3,14–16]. Rearranged NCCRs (rr-NCCR) display modified binding sites for cell-specific transcription factors, suggesting that NCCR rearrangements modify cellular tropism, and increase transcription and replication in CNS cells [1]. In addition, mutations affecting JCV binding to sialic acids have been evidenced in viral capsid protein VP1 sequences recovered from PML lesions [15,17].

Prototype JCV strains are known to differ from archetype strains. How and when JCV acquires neurotropic properties remains unknown. Few studies compared JCV genomes recovered from cerebral, blood and renal compartments [18–20,14,1]. Intra-compartment variability was analysed by sequencing cloned JCV NCCR and/or VP1 [3,15–17] or by deep sequencing techniques [7,21–23] confirming NCCR high variability, and in some cases suggesting the presence of NCCR and/or VP1 variants quasispecies [21–23]. To investigate emergence of neurotropic JCV, we deep sequenced JCV whole genome from CNS and urine samples of PML patients by single-molecule real-time (SMRT) sequencing [24], and studied inter host and intra-host variability.

Materials and Methods

Patients and samples

Viral DNA recovered from cerebrospinal fluid (CSF), cerebral biopsy (CB) and urine samples of patients diagnosed with PML between May 2011 and September 2018 was stored at -80°C in the virology laboratory of Cochin university hospital. Clinical samples, obtained to confirm JCV infection by standard virology diagnostic procedures, were stripped of personal identifiers other than sex, age and type of immunosuppression. The use of these samples for research ("Analysis of JCV NCCR in PML") was approved by the Ile de France I Ethics Committee (IRB: IORG0008367). Quantification was obtained by a real-time PCR specific for JCV T antigen coding sequence as previously described [1].

Full-length JCPyV genome amplification

After automated DNA extraction (QIAsymphony SP, QIAGEN), JCV full-length circular genome was amplified in two fragments overlapping at each extremity (Supplementary Figure 1). KAPA HiFi HotStart ReadyMix (Roche) was used to amplify 1 to 5 ng of DNA with 300 nM of primers (Supplementary Table 1) using a denaturation/activation step of 3 min at 95°C followed by 38 cycles of amplification (98°C for 20 s, 64°C or 66°C for fragments 1 and 2 respectively for 30 s, 72°C for 1 min 40 s). Amplified DNA was purified with magnetic beads Ampure PB (ref 100-265-900) and quantified using Qubit Fluorometer (Invitrogen, Life Technologies). DNA quality was checked using the Agilent 2100 Bioanalyzer (Agilent, Technologies).

SMRT Sequencing and Sequence Analysis

Barcoded libraries were prepared using the "Procedure & Checklist -Preparing Amplicon Libraries using PacBio® Barcoded Adapters for Multiplex SMRT® Sequencing" protocol (3.0 chemistry). Overlapping fragment pairs obtained from the same sample were tagged with the same barcode. All libraries were sequenced in one SMRTcell on the Sequel Sequencer. Each polymerase read is composed of several passes along the molecule. The different subreads were merged into Circular Consensus Sequence (CCS) reads. In order to get higher accuracy, CCS reads with at least 10 passes, and a length comprised between 2400 and 3400 nt were selected.

Data Analysis

The selected CCS reads from the sequencing step were analysed with a dedicated analysis pipeline available in the Sequana project [25], available on https://github.com/sequana/laa. The first step consisted in mapping the CCS reads on the archetype CY genomic sequence (Genbank: AB038249.1), which is based on the alignment program called minimap2 [26]. Then, consensus genomes were obtained for each barcode and variant calling performed using those consensus genomes.

JCV genotyping was performed by phylogenetic analysis of whole genome consensus sequences (NCCR excluded). Previously published JCV genomic sequences (n=113) from all known genotypes were obtained from the GenBank database and used as reference sequences for comparisons [27–34]. The phylogenetic analysis was decomposed into a multiple alignment with mafft followed by tree construction using RAxML. A standard phylogeny was repeated 100 times from which the best tree was selected. A bootstrap analysis was included, using 100 runs. Finally, iTOL (Tree of Life) web service was used for plotting (https://itol.embl.de/).

From the variant calling analysis, whole genome single mutations were filtered to keep single nucleotide variation (SNV) with a minimum depth of 10 and a minimum strand ratio of 0.2. No filtering was applied on the frequency of the SNVs so as to investigate low-frequency variation for each sample using the Freebayes variant calling software [35]. Missense mutations were analysed with regard to all samples of previous studies and NCBI database. Each NCCR sequence was aligned on the archetype CY NCCR (GenBank: AB038249.1) [9] with the BioEdit program and divided into sequence sections *a, b, c, d, e,* and *f* from 5' to 3' consisting of 25, 23, 55, 66, 18 and 69 bp respectively.

Results

JCV long-read sequencing

JCV whole genome from CNS and urine samples of PML patients (Table 1), was amplified in two overlapping fragments encompassing the complete NCCR and either the early or the late gene up to the inter region (Supplementary Figure 1). Using long read single molecule real time sequencing, high coverage was obtained (overlapping regions : mean 2661 +/- 1016; early and late coding region : mean 1481 +/- 698) and the average proportion of mapped reads was 98,2 +/- 2,6 % of the total reads. High fidelity CCS reads were obtained (average number of passes : 28, accuracy mean : 0.99956 corresponding to Phred quality score Q34). As expected, a uniform forward and reverse distribution was found with a 50/50 proportion on average.

JCV genotyping

Phylogenetic analysis was based on consensus sequences deduced from JCV whole genome except NCCR, compared to 113 reference JCV sequences from all known genotypes [27–34] (Figure 1). Viral strains distributed among 6 (1, 2, 3, 4, 6 and 7) out of 7 JCV genotypes, each associated with a geographic human population [36], reflecting the known distribution of genotypes in European countries. Importantly, each patient harbored the same and single genotype in both urine and cerebral compartments.

JCV genotyping was previously based on distribution of single nucleotide polymorphisms (SNP) over the whole coding region [36,37]. We calculated the error rate of SMRT sequencing technique by analyzing variability at 55 SNP positions. Among 46 samples, 176/1 246 595 CCS reads harbored mutations, corresponding to a 0,014% error rate. We therefore considered significant point mutations if detected in ≥ 0,5% reads.

NCCR variability

In NCCR sequences from CNS samples, various combinations of deletions/insertions defined 40 unique rearranged profiles. Deletions (median size 48 bp) found in 24/24 cases mostly affected section *d* (18/24 cases). In 15 cases, deletion breakpoints clustered in a 10 nucleotide sequence close to *c-d* junction (Figure 2). Other large deletions affected sections *e* and *f*, less often *b*. In 21/24 cases, duplicated fragments (median size 83 bp), mostly from sections *b-c,* were inserted near the *e-f* junction (n=15) or in sections *c, d,* and at the 5' extremity of the Late coding region (Agnoprotein coding sequence) (Figure 3). In one case a 20 bp duplicated fragment from *vp1* gene (nt 1860-1879, numbering in reference to archetype CY (AB038249.1)) was inserted in NCCR (Supplementary Figure 2). In 23 cases, rr-NCCR exhibited a deleted sequence and a duplicated fragment which encompassed the exactly similar deleted sequence (Supplementary Table 2 and Supplementary Figure 3). Most importantly, among all NCCR profiles from brain-derived samples, no archetypal sequence was observed.

Intra-compartment variability was analyzed in 23 CNS samples showing sufficient coverage (> 1000 CCS reads). A single NCCR variant was identified in 14/23 samples, whereas 9 samples contained

two, three or more NCCR variants (in 5, 2 and 2 cases respectively). In 7 / 9 cases, the mixed population consisted of a highly predominant variant (> 90%), associated with minor subpopulations (Supplementary Table 2). Importantly, in each sample, multiple rr-NCCR variants shared at least one rearrangement, consisting of a deletion in all cases except one. Mixed NCCR populations suggested sequential (Figure 4) or divergent (Supplementary Table 2) evolution from a primary variant. One patient had two CSF samples drawn at a 19 day interval, and presented an additional NCCR variant in the second sample (Supplementary Table 2). In one sample, two major NCCR variants shared three identical rearrangements, and multiple minor variants exhibited distinct combinations of deletions grouped in a short sequence (Figure 4). Deletions had sequential increasing sizes from 3 to 47 bp and their breakpoints were located at every nucleotide position of section *f* from nt 206 to nt 254 (Supplementary Figure 4 and Supplementary Table 3).

NCCR recovered from urine samples showed a 100% archetypal structure in most PML and all control cases. Short insertions/deletions (2-38 nt) were however found in either the major or minor viral subpopulations of 3 and 5 urine samples from PML patients respectively (Supplementary Table 2). In 3 patients, rr-NCCR differed in CSF samples and in urine obtained 18, 7 and 54 days later respectively. Moreover, in one patient, rr-NCCR displayed opposite patterns of deletions and duplications in both compartments, as shown by suppression of binding sites for transcription factors Sp1 and NFI in the CSF rr-NCCR, contrasting with duplication of these binding sites in urine rr-NCCR (Figure 4). Finally, one urine sample contained multiple NCCR subpopulations (Supplementary Figure 5), characterized by insertions of short duplicated fragments originating from a 23 nt sequence of the *d-e* junction and showing sequentially increasing sizes from 3 to 15 nt (Supplementary Table 4).

Coding regions

VP1 is the major JCV capsid protein implicated in binding to cells, and several VP1 mutations distinctive of PML have been described at amino acid (aa) positions 50, 51, 55, 60, 61, 122, 129, 223, 265, 267, 269 and 271. In CNS samples, several mutations resulted in 2 to 3 known or novel aminoacid substitutions at 7 of these 12 positions (Supplementary Table S2). Ser 269 was most frequently altered (n=10), followed by Leu 55 (n=6) and Asn 265 (n=4). Novel mutations affecting Phe68 and Tyr81 were detected in two samples.

Intra-compartment analysis showed that 4/ 23 samples contained a 100% wild type (wt) VP1 population. In 10 samples, the whole population had a single mutation affecting aa positions 265, 267 or 269 in 7/10 cases, and aa 55 or 61 in 3/10 cases. Interestingly, 10 other samples contained mixed subpopulations, 9 of which associated wt VP1 with single-mutated and double-mutated VP1 molecules, the latter representing <5% of the viral population (Table 2).

Other JCV coding regions were analyzed. While most reported missense mutations in VP2 minor capsid protein gene reflected genotype specificity [7,15,22,23], we found novel genotype-independent mutations (A20T, A30V, A36T and P65A) in 4/24 samples (Supplementary Table 2). In addition, three missense mutations (Y407N, M402I, G255K) affected the helicase domain, and one affected the Pol alpha binding and DNA binding (Zn finger) domain (N299T) of LT protein involved in viral replication.

Importantly, whereas no CNS sample contained archetype NCCR, most showed wt VP1. In most samples containing multiple NCCR variants showing sequential rearrangements, each NCCR variant harboured either wt VP1, or single or double-mutated VP1. Importantly, VP1 mutations were distributed independently from the NCCR rearrangement (Supplementary Table 5), as exemplified in particular in one sample containing a rr-NCCR variant preferentially showing a L55F-mutated VP1, while a second rr-NCCR variant was mainly associated with a S269F-mutated VP1 (Table 3).

Discussion

PML occurs after deep and sustained immunosuppression mostly secondary to HIV infection, hemopathies, or immunosuppressive therapies. Rearrangements in NCCR and mutations in VP1 are a hallmark of viral genomes recovered from the brain of PML patients, but how and when these mutations originate is not known. Deep sequencing technologies detect minor viral variants and describe the dynamics of viral populations [7,21–23]. However, technologies based on fragmentation of amplified products are inappropriate to identify NCCR large deletions or duplications and to associate mutations on distant portions of viral genomes. Errors generated by PacBio SMRT technology for long fragment sequencing (> 20 kbases) [24] are randomly distributed and mainly consist of single insertions/deletions, or more rarely mismatches [38]. By using filtering criteria set up to select high fidelity reads with a number of passes $\geq$ 10 and minimum accuracy of 0.999, we obtained high coverage of JCV whole genome and appropriate depth to analyse missense mutated subpopulations $\geq$ 0,5% while NCCR rearrangements characterized by insertion/deletions ≥ 2 nt were detected with ≥ 0,1% depth.

PML strains showed no preferential genotypic distribution other than local epidemiology [31]. Each patient was infected with a single and same genotype in the CNS and urine, supporting the hypothese that co-infection or reinfection is a rare event.

Our results showed a common pattern of NCCR rearrangements suggesting a common, albeit as yet unknown, mechanism. NCCR rearrangements are thought to result from recombination events during high replication levels permitted by pathological suppression of immune control. Deletions of section *d* have frequently been described in PML cerebral strains [15], and breakpoint profiles have been reported [39]. We identified a hotspot of deletion breakpoints concentrated along a short 10 nt sequence which superimposes with a palindromic sequence described from nt 110 to 131, and which could lead to a stable hairpin DNA secondary structure [40]. This conformation could favor replication fork stalling or pausing, resulting in double strand DNA breaks requiring mechanisms of reparation involving recombination (homologous recombination and nonhomologous end joining). In our study, *f* section was frequently altered. Although not initially included in the classification of JCV variants scheme, deletions of this section have subsequently been described in several patients, mostly combined with duplications and rarely as a single rearrangement [1]. Moreover, we observed duplicated fragments inserted in the 5' extremity of Agnoprotein gene which have very rarely been described [41,42]. Finally, a NCCR rearrangement involved a fragment from the *vp1* gene, 1681 bp distant from insertion site. Recently, Hu *et al* [43] described insertion, in *f* section, of a 65 bp fragment from LTAg (polymerase alpha binding site), suggesting that NCCR rearrangements may involve duplications and insertions from distant genomic regions.

Our results showed that in immunosuppressed PML patients, high uncontrolled viral replication levels could generate minor rearranged NCCR populations in the renal compartment, evolving independently from brain-derived viral populations, as suggested by divergent patterns observed in both compartments, even at short intervals. These results differ from previous studies showing that urine harbored archetypal NCCR in 15/18 patients [14,15,18–21,44,45], or similar variations in urine and CNS in 3 cases [18,44].

Binding sites for transcription factors have been identified in NCCR and we previously showed that a single deletion in section *f* modifies NCCR-driven tissue-specific expression [1]. By showing that deletions in *d* section resulted in suppression of binding sites for transcription factors SP1 and NFI [46], and by evidencing opposite patterns of suppression or duplication of these binding sites in variants recovered from urine and CNS, our results strongly reinforce the tissue-specific adaptive role of NCCR rearrangements.

In two cases, multiple minor rr-NCCR populations carried various combinations of insertions/deletions concentrated in a short sequence. In each case, the nature and location of NCCR rearrangements (deletions in *f* section, and duplications from *d* and *e* sections respectively) were consistent with rearrangements observed in the respective CSF and urinary compartments of PML patients. However, the serial location of breakpoints and the sequential increasing sizes of insertions/deletions suggested a random recombination mechanism requiring further investigation.

Mutations in *vp1* caused amino acid substitutions in sialic acid binding regions and in antigenic epitopes, as most frequently described [7,15,17,22,23]. Two newly identified amino acid substitutions (F68L and Y81C) are located in one of the three most external loops (BC) of VP1 and could impact its conformation and tissue tropism by modifying sialic acid receptor binding [17] and/or participate in immune response evasion [47]. Finally, VP2 and VP3 minor coats proteins are necessary for efficient viral propagation and LT holds main functions necessary to viral replication. However, the impact of newly described mutations in VP2 N-terminal domain and in the helicase domain of LT will necessitate further studies.

We were able to deduce the chronology of emergence of viral populations by studying inter- and intra-compartment variability and by analyzing the combination of NCCR and VP1 mutations at the single molecule level. First, the absence of archetypal NCCR sequence in brain-derived samples favors the hypothesis that initial NCCR rearrangements might occur in a peripheral compartment. Similar results have been described in a study of 3 cases using the same technology [22]. Archetype JCV has very rarely (5/261 cases) been detected in the CSF of PML patients [21,45]. Second, when multiple NCCR populations were observed in a sample, they shared one or more identical rearrangements, suggesting they stemmed from an initial variant, sufficient to drive replication and dissemination in the cerebral tissue, and which subsequently continued to evolve. Interestingly, the common rearrangement consisted in a deletion in all cases except one, favoring the hypothesis that a deletion in NCCR is an initial predominant step, followed by further deletions, duplications and insertions in this region [40]. In most samples containing multiple rearranged forms, one major NCCR form widely predominated as previously observed with an equivalent sequencing depth in 8/10 CSF samples, all showing minor variants [21]. Moreover, diversification of NCCR populations was observed in a shorter interval than previously reported [16]. Finally, whereas all CNS viral populations showed rr-NCCR, half of samples contained either only wt VP1, or wt VP1

subpopulations. This suggests that NCCR rearrangements are required for cerebral infection [14,48] whereas VP1 mutations occur later, when infection progresses [49]. In particular, some samples carried three different rr-NCCR variants, possibly attesting for a longer duration of evolution. Each rr-NCCR variant comprised subpopulations still harbouring wt VP1, while other subpopulations had adopted one or two secondary mutations in VP1 on different key positions of the protein, suggesting that VP1 mutations are later events not required for viral replication in brain tissue**.** A progressive apparition of VP1 mutation in the course of infection has been observed in a humanized model of myelin-deficient mice with humanized white matter [50]. Mutations at positions 55, 60, 265, 267 and 269 of VP1 were predominantly mutually exclusive, suggesting they confer a sufficient benefit for viral tropism and replication [17]. However, minor viral populations carried multiple combinations of two mutations, suggesting a constant evolution during viral adaptation.

In conclusion, whole genome SMRT sequencing of JCV in PML further illustrates wide intra-host genomic variability and the existence of quasispecies, which parallels previous observations in RNA viruses [7,21,23]. Deep analysis of the structure of NCCR variants demonstrated common patterns, and suggested that neurotropic variants stem from a unique rearranged variant undergoing further evolution. By shedding light on emergence of viral subpopulations, our results provide new insight in the genesis of JCV neurotropic strains and PML pathogenesis.

Funding : this work was financed by Assistance Publique-Hôpitaux de Paris (APHP) and Institut National de la Santé et de la Recherche Médicale (INSERM)

Conflict of Interests : the authors declare no conflict of interest

References

1. L'Honneur A-S, Leh H, Laurent-Tchenio F, Hazan U, Rozenberg F, Bury-Moné S. Exploring the role of NCCR variation on JC polyomavirus expression from dual reporter minicircles. PLoS One. **2018**; 13(6):e0199171.

2. Kitamura T, Aso Y, Kuniyoshi N, Hara K, Yogo Y. High incidence of urinary JC virus excretion in nonimmunosuppressed older patients. J Infect Dis. **1990**; 161(6):1128–1133.

3. Tan CS, Ellis LC, Wüthrich C, et al. JC virus latency in the brain and extraneural organs of patients with and without progressive multifocal leukoencephalopathy. J Virol. **2010**; 84(18):9200–9209.

4. Kato A, Kitamura T, Takasaka T, et al. Detection of the archetypal regulatory region of JC virus from the tonsil tissue of patients with tonsillitis and tonsilar hypertrophy. J Neurovirol. **2004**; 10(4):244–249.

5. Berger JR, Miller CS, Danaher RJ, et al. Distribution and Quantity of Sites of John Cunningham Virus Persistence in Immunologically Healthy Patients: Correlation With John Cunningham Virus Antibody and Urine John Cunningham Virus DNA. JAMA Neurol. **2017**; 74(4):437–444.

6. Bayliss J, Karasoulos T, McLean CA. Immunosuppression Increases JC Polyomavirus Large T Antigen DNA Load in the Brains of Patients Without Progressive Multifocal Leukoencephalopathy. J Infect Dis. **2013**; 207(1):133–136.

7. Chalkias S, Gorham JM, Mazaika E, et al. ViroFind: A novel target-enrichment deep-sequencing platform reveals a complex JC virus population in the brain of PML patients. PLoS One. **2018**; 13(1):e0186945.

8. Major EO, Yousry TA, Clifford DB. Pathogenesis of progressive multifocal leukoencephalopathy and risks associated with treatments for multiple sclerosis: a decade of lessons learned. Lancet Neurol. **2018**; 17(5):467–480.

9. Yogo Y, Kitamura T, Sugimoto C, et al. Isolation of a possible archetypal JC virus DNA sequence from nonimmunocompromised individuals. J Virol. **1990**; 64(6):3139–3143.

10. Jensen PN, Major EO. A classification scheme for human polyomavirus JCV variants based on the nucleotide sequence of the noncoding regulatory region. J Neurovirol. **2001**; 7(4):280–287.

11. Egli A, Infanti L, Dumoulin A, et al. Prevalence of polyomavirus BK and JC infection and replication in 400 healthy blood donors. J Infect Dis. **2009**; 199(6):837–846.

12. Bofill-Mas S, Girones R. Excretion and transmission of JCV in human populations. J Neurovirol. **2001**; 7(4):345–349.

13. Frisque RJ, Bream GL, Cannella MT. Human polyomavirus JC virus genome. J Virol. **1984**; 51(2):458–469.

14. Gosert R, Kardas P, Major EO, Hirsch HH. Rearranged JC virus noncoding control regions found in progressive multifocal leukoencephalopathy patient samples increase virus early gene expression and replication rate. J Virol. **2010**; 84(20):10448–10456.

15. Reid CE, Li H, Sur G, et al. Sequencing and analysis of JC virus DNA from natalizumab-treated PML patients. J Infect Dis. **2011**; 204(2):237–244.

16. Nakamichi K, Kishida S, Tanaka K, et al. Sequential changes in the non-coding control region sequences of JC polyomaviruses from the cerebrospinal fluid of patients with progressive multifocal leukoencephalopathy. Arch Virol. **2013**; 158(3):639–650.

17. Gorelik L, Reid C, Testa M, et al. Progressive multifocal leukoencephalopathy (PML) development is associated with mutations in JC virus capsid protein VP1 that change its receptor specificity. J Infect Dis. **2011**; 204(1):103–114.

18. Vaz B, Cinque P, Pickhardt M, Weber T. Analysis of the transcriptional control region in progressive multifocal leukoencephalopathy. J Neurovirol. **2000**; 6(5):398–409.

19. Han GP, Miura K, Ide Y, Tsutsui Y. Genetic analysis of JC virus and BK virus from a patient with progressive multifocal leukoencephalopathy with hyper IgM syndrome. J Med Virol. **2005**; 76(3):398–405.

20. Delbue S, Sotgiu G, Fumagalli D, et al. A case of a progressive multifocal leukoencephalopathy patient with four different JC virus transcriptional control region rearrangements in cerebrospinal fluid, blood, serum, and urine. J Neurovirol. **2005**; 11(1):51–57.

21. Van Loy T, Thys K, Ryschkewitsch C, et al. JC virus quasispecies analysis reveals a complex viral population underlying progressive multifocal leukoencephalopathy and supports viral dissemination via the hematogenous route. J Virol. **2015**; 89(2):1340–1347.

22. Seppälä H, Virtanen E, Saarela M, et al. Single-Molecule Sequencing Revealing the Presence of Distinct JC Polyomavirus Populations in Patients With Progressive Multifocal Leukoencephalopathy. J Infect Dis. **2016**; .

23. Takahashi K, Sekizuka T, Fukumoto H, et al. Deep-Sequence Identification and Role in Virus Replication of a JC Virus Quasispecies in Patients with Progressive Multifocal Leukoencephalopathy. J Virol. **2017**; 91(1).

24. Nakano K, Shiroma A, Shimoji M, et al. Advantages of genome sequencing by long-read sequencer using SMRT technology in medical area. Hum Cell. **2017**; .

25. Cokelaer T, Desvillechabrol D, Legendre R, Cardon M. "Sequana": a Set of Snakemake NGS pipelines. JOSS. **2017**; 2(16):352.

26. Li H. Minimap2: pairwise alignment for nucleotide sequences. Birol I, editor. Bioinformatics. **2018**; 34(18):3094–3100.

27. Jobes DV, Chima SC, Ryschkewitsch CF, Stoner GL. Phylogenetic analysis of 22 complete genomes of the human polyomavirus JC virus. J Gen Virol. **1998**; 79 (Pt 10):2491–2498.

28. Schaffer K, Sheehy N, Coughlan S, Bergin C, Hall WW. JC virus in the Irish population: significant increase of genotype 2 in immunocompromised individuals. J Neurovirol. **2006**; 12(1):39–46.

29. Agostini HT, Deckhut A, Jobes DV, et al. Genotypes of JC virus in East, Central and Southwest Europe. J Gen Virol. **2001**; 82(Pt 5):1221–1331.

30. Cui X, Wang JC, Deckhut A, et al. Chinese strains (Type 7) of JC virus are afro-asiatic in origin but are phylogenetically distinct from the Mongolian and Indian strains (Type 2D) and the Korean and Japanese strains (Type 2A). J Mol Evol. **2004**; 58(5):568–583.

31. Fink MCD, Oliveira ACP de, Romano CM, et al. Molecular characterization of human polyomavirus JC in Brazilian AIDS patients with and without progressive multifocal leukoencephalopathy. J Clin Virol. **2010**; 48(1):6–10.

32. Karalic D, Lazarevic I, Knezevic A, Cupic M, Jevtovic D, Jovanovic T. Distribution of JC virus genotypes among Serbian patients infected with HIV and in healthy donors. J Med Virol. **2014**; 86(3):411–418.

33. Comerlato J, Campos FS, Oliveira MT, et al. Molecular detection and characterization of BK and JC polyomaviruses in urine samples of renal transplant patients in Southern Brazil. J Med Virol. **2015**; 87(3):522–528.

34. Hu C, Huang Y, Su J, Wang M, Zhou Q, Zhu B. Detection and analysis of variants of JC polyomavirus in urine samples from HIV-1-infected patients in China's Zhejiang Province. J Int Med Res. **2018**; :300060517746297.

35. Garrison E, Marth G. Haplotype-based variant detection from short-read sequencing. arXiv:12073907 [q-bio] [Internet]. **2012** [cited 2019 Feb 13]; . Available from: http://arxiv.org/abs/1207.3907

36. Cubitt CL, Cui X, Agostini HT, et al. Predicted amino acid sequences for 100 JCV strains. J Neurovirol. **2001**; 7(4):339–344.

37. Stoner GL, Jobes DV, Fernandez Cobo M, Agostini HT, Chima SC, Ryschkewitsch CF. JC virus as a marker of human migration to the Americas. Microbes Infect. **2000**; 2(15):1905–1911.

38. Wenger AM, Peluso P, Rowell WJ, et al. Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome. Nat Biotechnol. **2019**; 37(10):1155–1162.

39. Agostini HT, Ryschkewitsch CF, Singer EJ, Stoner GL. JC virus regulatory region rearrangements and genotypes in progressive multifocal leukoencephalopathy: two independent aspects of virus variation. J Gen Virol. **1997**; 78 (Pt 3):659–664.

40. Johnson EM, Wortman MJ, Dagdanova AV, Lundberg PS, Daniel DC. Polyomavirus JC in the context of immunosuppression: a series of adaptive, DNA replication-driven recombination events in the development of progressive multifocal leukoencephalopathy. Clin Dev Immunol. **2013**; 2013:197807.

41. Yogo Y, Kitamura T, Sugimoto C, et al. Sequence rearrangement in JC virus DNAs molecularly cloned from immunosuppressed renal transplant patients. J Virol. **1991**; 65(5):2422–2428.

42. Yasuda Y, Yabe H, Inoue H, et al. Comparison of PCR-amplified JC virus control region sequences from multiple brain regions in PML. Neurology. **2003**; 61(11):1617–1619.

43. Hu C-Q, Su J-W, Wang M-Y, et al. Sequencing and analysis of John Cunningham polyomavirus DNA from acquired immunodeficiency syndrome patients with progressive multifocal leukoencephalopathy. Chin Med J. **2020**; 133(23):2887–2889.

44. Pfister LA, Letvin NL, Koralnik IJ. JC virus regulatory region tandem repeats in plasma and central nervous system isolates correlate with poor clinical outcome in patients with progressive multifocal leukoencephalopathy. J Virol. **2001**; 75(12):5672–5676.

45. Domínguez-Mozo MI, García-Montojo M, Arias-Leal A, et al. Monitoring the John Cunningham virus throughout natalizumab treatment in multiple sclerosis patients. Eur J Neurol. **2016**; 23(1):182–189.

46. Kim J, Woolridge S, Biffi R, et al. Members of the AP-1 family, c-Jun and c-Fos, functionally interact with JC virus early regulatory protein large T antigen. J Virol. **2003**; 77(9):5241–5252.

47. Ray U, Cinque P, Gerevini S, et al. JC polyomavirus mutants escape antibody-mediated neutralization. Sci Transl Med. **2015**; 7(306):306ra151.

48. Marshall LJ, Ferenczy MW, Daley EL, Jensen PN, Ryschkewitsch CF, Major EO. Lymphocyte gene expression and JC virus noncoding control region sequences are linked with the risk of progressive multifocal leukoencephalopathy. J Virol. **2014**; 88(9):5177–5183.

49. Wharton KA, Quigley C, Themeles M, et al. JC Polyomavirus Abundance and Distribution in Progressive Multifocal Leukoencephalopathy (PML) Brain Tissue Implicates Myelin Sheath in Intracerebral Dissemination of Infection. Major EO, editor. PLoS One. **2016**; 11(5):e0155897.

50. Kondo Y, Windrem MS, Zou L, et al. Human glial chimeric mice reveal astrocytic dependence of JC virus infection. J Clin Invest. **2014**; 124(12):5323–5336.

Figure captions (as a list)

Figure 1. Circular phylogenetic tree of JCV strains from PML and control patients.

Maximum-likelihood (ML) phylogenetic tree generated by PhyML software, based on whole genome alignment (with the exception of NCCR) from 46 samples and 113 reference sequences obtained from GenBank (listed in Materials and Methods Section). Cases 1 to 32, PML patients, cases 33 to 37, controls. C: CSF, U: Urine, CB: Cerebral biopsy.

Figure 2. Patterns of JCV NCCR deletions in PML CNS samples.

Fourty NCCR deletion patterns were found in 24 CNS samples. Deletions are represented with their size and breakpoint origin along the reference archetypal "CY" NCCR sequence (AB038249.1); numbering as in Yogo et al, 1990. Deletions involve mainly the d and f sections. A cluster (nt 109-119) of deletion breakpoints is enlarged.

Figure 3. Patterns of JCV NCCR duplications/insertions in PML CNS samples.

Twenty-six main NCCR duplication/insertion patterns were found in 21 CNS samples. A) Site of insertion and size of the duplicated fragment : as in Figure 2. A cluster of insertion sites at e-f junction is enlarged. B) Origin of duplicated fragments, mostly affecting section b.

Figure 4. Compartmentalization of rr-NCCR.

NCCR sequences from patient's 5 CSF and urine are aligned along the reference archetype as in Figure 2. Deletions : brackets; duplications : grey bars. Horizontal dotted line : multiple mixed subpopulations carrying combinations of small deletions in a short region of section f. AP-1, Sp1, NFI, Spi-B and p53 binding sites are represented below the reference sequence. Divergent patterns in both compartments suggest tissue specificity.

Table(s) with caption(s) (on individual pages);

Patient	Sex	Age (y)	Underlying disease	Sample	JCV viral load (log copy/mL)	NCCR sequence coverage
1	M	53	HIV	CSF	6,4	4217
				Urine° (18)	7,3	2894
2	M	42	HIV	CSF	4,7	1289
				Urine° (13)	6,9	3018
3	M	84	CLL	CSF	4,4	2285
				Urine° (7)	5,4	1602
4	M	57	HIV	CSF	6,4	3519
				Urine° (33)	9,2	56
5	M	62	HIV	CSF	4,5	3670
				Urine° (54)	7,5	2615
6	F	41	HIV	CSF	4,3	3919
				Urine° (13)	7,6	2442
7	M	41	HIV	CSF	4,5	2127
				Urine° (6)	5,9	2326
8	M	60	HIV	CSF	4,0	2698
				Urine° (26)	8,5	1605
9	M	49	MZL	CSF n°1	4,2	4441
				CSF n°2	5,0	4114
10	F	51	Lung transplant	CSF	3,6	4194
11	M	45	NA	CSF	5,1	1367
12	F	69	HIV	CSF	6,3	1529

13	M	19	PID (MST1 deficiency) / HL	CSF	6,4	2345
14	M	40	HIV	CSF	3,6	4046
15	M	45	HIV	CSF	5,8	2557
16	F	70	unknown immunodeficiency	CSF	6,7	2951
17	M	27	HIV	CSF	7,3	88
18	F	42	HIV	CSF	5,9	3552
19	M	55	LPD	CSF	4,2	2250
20	F	23	SCID	CSF	6,6	1795
21	M	19	WAS	CSF	4,9	2673
22	M	33	NA	CSF	2,4	2162
23	F	40	HIV	CSF	6,7	3478
24	F	46	MS/NTZ	CB	NA	3856
25	F	54	HIV	Urine	5,9	2078
26	F	48	HIV	Urine	4,5	1846
27	F	45	NA	Urine	6,2	2353
28	F	51	HIV	Urine	8,7	2668
29	M	91	lymphopenia of unknown origin	Urine	7,5	2710
30	M	47	HIV	Urine	4,7	1248
31	F	66	MS/DMF	Urine	4,5	3348
32	M	24	CID of unknown origin	Urine	5,7	3256
33*	M	67	NA	Urine	5,8	3142
34*	M	34	MS	Urine	8,4	3462
35*	M	84	leukemia	Urine	3,7	2188
36*	F	43	lymphoma	Urine	4,5	3136
37*	F	11	histiocytosis	Urine	7,1	3290

Table 1. Characteristics of patients and samples

*: Non PML control patients; ◦ delay after CSF sample (days); M: male; F: female; BMT: Bone Marrow Transplant; CB: cerebral biopsy; CID : Combined

Immunodeficiency ; CLL: chronic lymphocytic leukemia; CSF: cerebrospinal fluid; HL: Hodgkin Lymphoma; LPD: Lymphoproliferative Disorder; MS/DMF: Multiple sclerosis treated with dimethylfumarate. MS/NTZ: Multiple sclerosis treated with Natalizumab; MZL: Marginal Zone Lymphoma; NA: Not available; PID: Primary Immunodeficiency ; SCID : Severe combined immunodeficiency ; WAS: Wiskott–Aldrich syndrome.

Patient no.	VP1 coverage	Frequency (%)	Leu 55	Lys 60	His 122	Asn 265	Ser 267	Ser 269	Gln271
1	2373								
		56,4	Phe	-		-	-	-	-
		34,2	-	-		-	-	-	-
		5,8	-	Thr		-	-	-	-
		3,7	-	-		-	-	Phe	-
2	1261								
		91,9	Phe	-		-	-	-	-
		4,5	-	-		-	-	Phe	
		2,1	Phe	-		-	-	Tyr	-
		1,5	-	-		-	-	-	-
3	1254								
		66,3	-	-		Asp	-	-	-
		22,3	-	-		-	-	Phe	-
		5,3	-	-		-	-	-	-
		4,7	Phe	-		-	-	-	-
		1,4	-	-		Asp	-	Phe	-
10	4224								
		97,4					Ile		
		2,6			Tyr		Ile		
11	1805								
		52,4	-	Asn		-	-	-	-
		26,3	His	-		-	-	-	-
		7,6	-	-		-	-	-	-
		7,5	-	-		-	Cys	-	-
		4,2	His	Asn		-	-	-	-
		1,9	-	Asn		-	Cys	-	-
12	1356								

		50,4	-	-	-	-	Tyr	-
		42,4	-	-	His	-	-	-
		4,4	-	-	His	-	Tyr	-
		2,8	-	-	-	-	-	-
14	2885							
		88,4	-	-	-	-	Ala	-
		8,1	-	-	-	-	-	His
		3,1	-	-	-	-	-	-
		0,5	-	-	-	-	Ala	His
21	1875							
		67,3	Phe	-	-	-	-	-
		23,0	-	-	-	-	-	-
		9,1	-	-	-	-	Phe	-
		0,6	Phe	-	-	-	Phe	-
22	1223							
		95,9	-	-	-	Ala	-	
		4,1	-	-	-	-	-	
24	3982							
		63,3	-	-	-	-	-	-
		36,8	-	-	-	-	Phe	-

Table 2. Distribution of VP1 subpopulations in PML CNS samples.

Among 10 samples containing mixed VP1 subpopulations, 9 associate various proportions of wild type, single and/or double mutated VP1.

NCCR		VP1			Total No. of reads
del 190-255 (66 bp)	del 190-255 (66 bp) ins 271 (153-189//256-271)	1622C>T	2265C>A	2265C>T	
19	3	-	-	-	22
1159	2	L55F	-	-	1161
3	52	-	-	S269F	55
23	-	L55F	S269Y	-	23
3	-	-	S269Y	-	3

Table 3. Emergence of VP1 mutations in rearranged NCCR variants.

In one CSF (patient 2), two NCCR variants sharing a common deletion acquire mutated VP1, each preferentially associated with a distinct VP1 mutation. Each line indicates the number of reads corresponding to each subpopulation defined by NCCR variation and VP1 mutation. NCCR variants are depicted in Supplementary Figure S3.

Figure 1

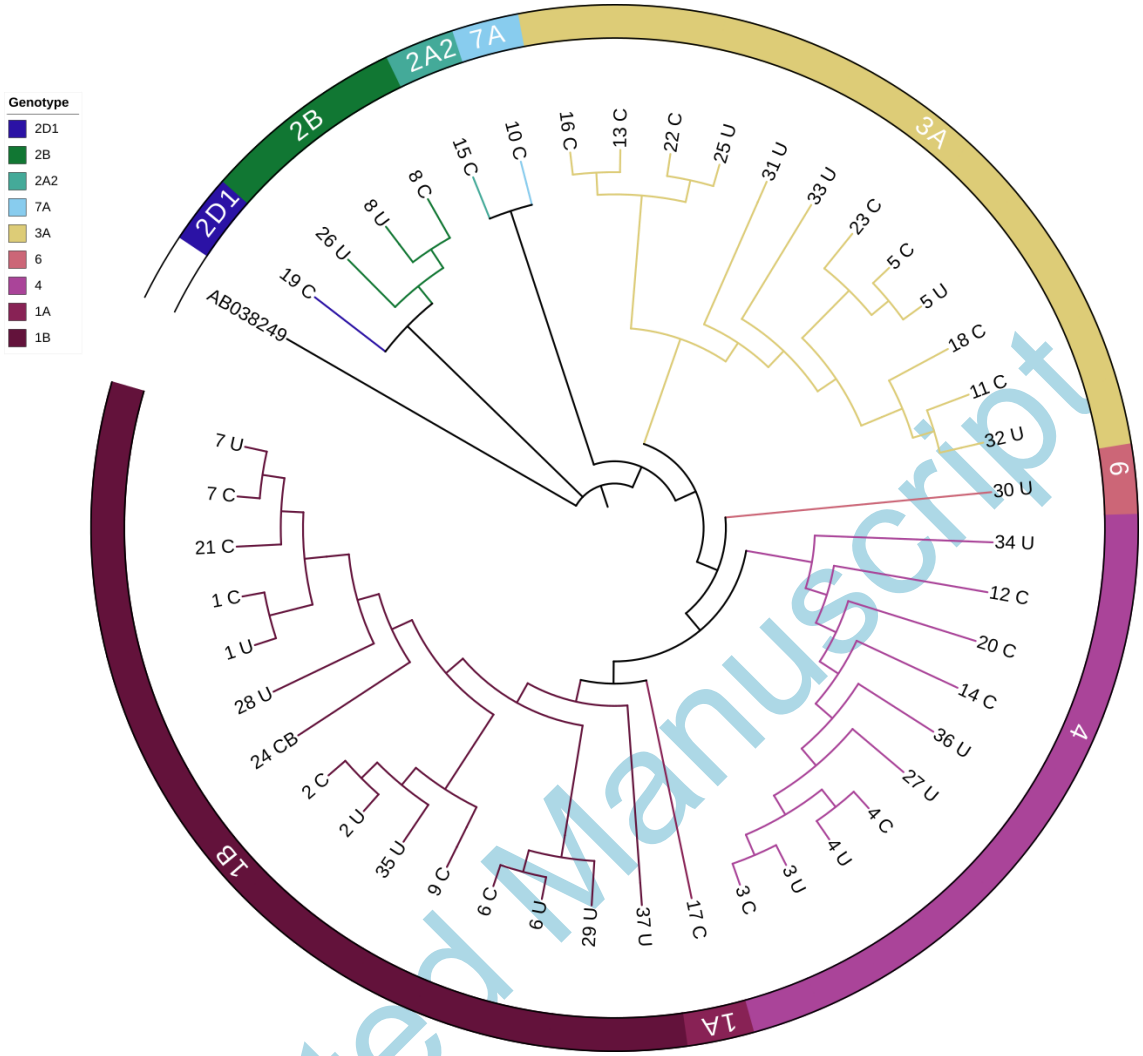

Figure 2

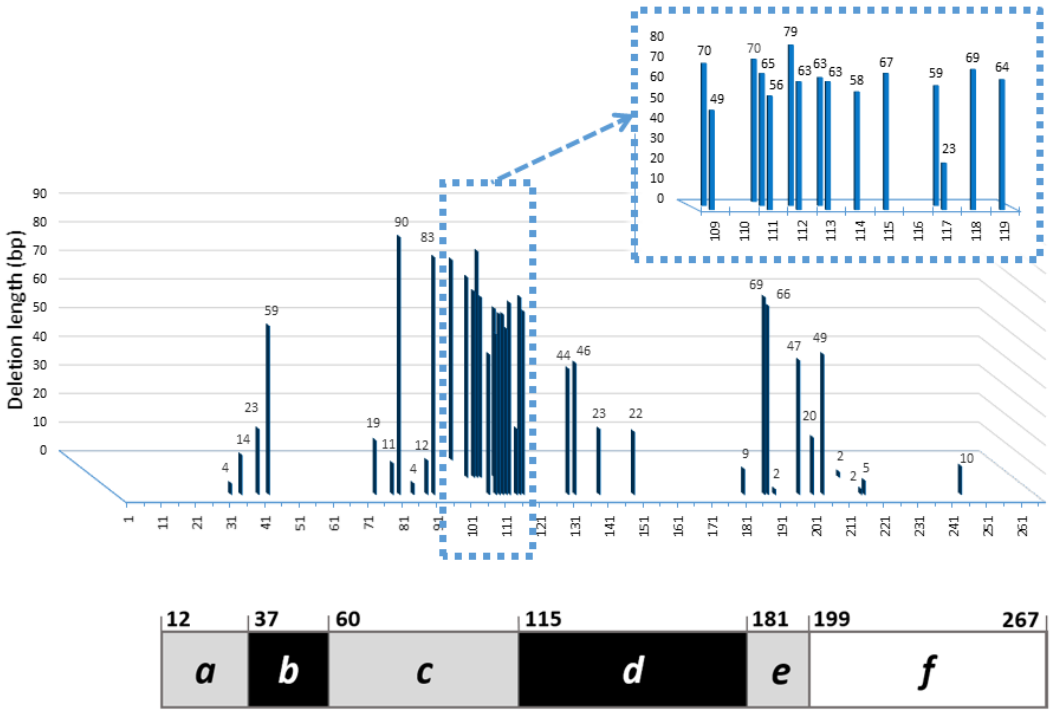

Figure 3

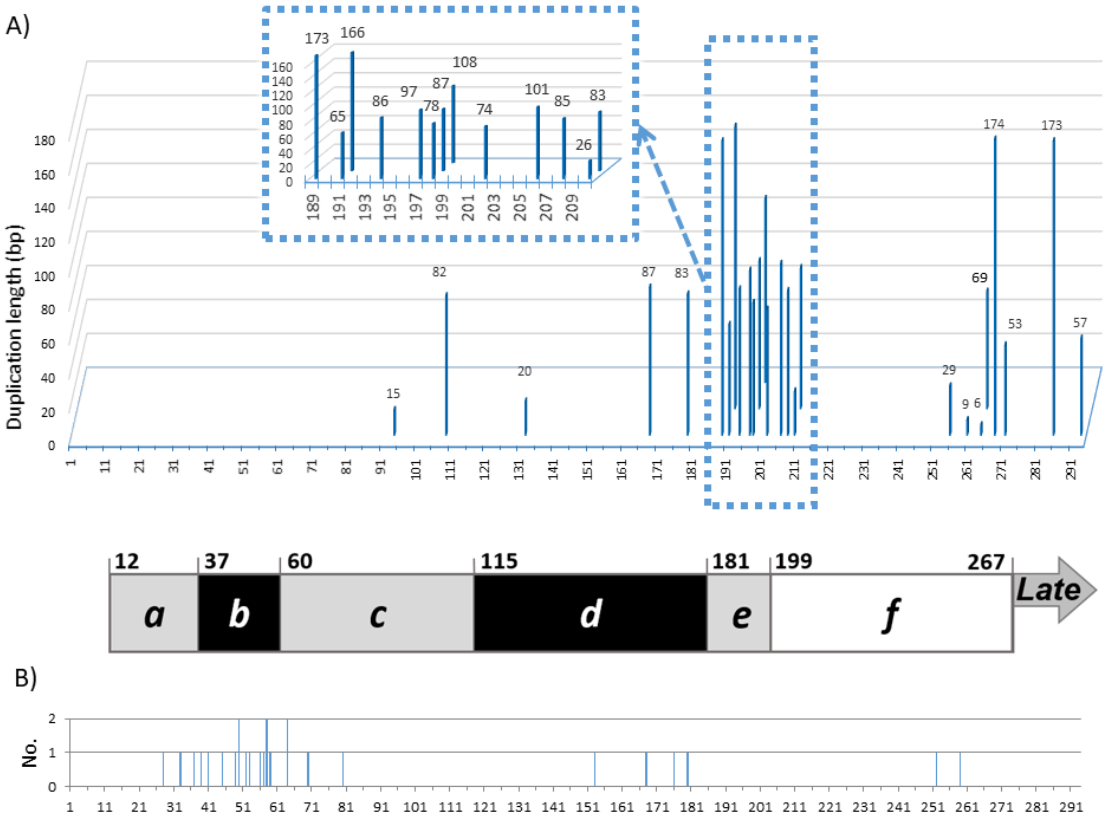

Figure 4