Nucleocytoviricota viral factories are transient organelles made by phase separation

Sofia Rigou[1]#, Alain Schmitt[1]#, Audrey Lartigue[1], Lucile Danner[1], Claire Giry[1], Feres Trabelsi[1], Lucid Belmudes[2], Natalia Olivero-Deibe[3], Yohann Couté[2], Mabel Berois[4], Matthieu Legendre[1], Sandra Jeudy[1], Chantal Abergel[1]*, Hugo Bisio[1]*.

[1] Aix-Marseille University, Centre National de la Recherche Scientifique, Information Génomique & Structurale (IGS), Unité Mixte de Recherche 7256 (Institut de Microbiologie de la Méditerranée, FR3479), IM2B, IOM, 13288 Marseille Cedex 9, France.

[2] Univ. Grenoble Alpes, INSERM, CEA, UA13 BGE, CNRS, CEA, FR2048, 38000 Grenoble, France.

[3] Immunovirology Lab, Institut Pasteur de Montevideo, 11400, Montevideo, Uruguay

[4] Sección Virología, Instituto de Química Biológica, Facultad de Ciencias, Universidad de la República, Montevideo 11600, Uruguay

#Co-first Author

* Correspondence: hugo.bisio@igs.cnrs-mrs.fr, chantal.abergel@igs.cnrs-mrs.fr

Phase separation is a common mechanism utilized by viruses to achieve replication, host manipulation and virion morphogenesis. The newly defined phylum *Nucleocytoviricota* encompass ubiquitous and diverse viruses including *Poxviridae,* the climate-modulating *Emiliania huxleyi* virus and the previously termed Nucleocytoplasmic large DNA viruses (NCLDV). Cytoplasmic members of this phylum form viral factories but their nature remains unknow. Here we show that these viral factories are formed by phase separation. We demonstrate that mimivirus viral factories are formed by multilayered phase separation using at least two scaffold proteins. We also generate a pipeline to bioinformatically identify putative scaffold proteins in all other *Nucleocytoviricota* despite major primary sequence variability. Such predictions were based on a conserved molecular grammar governed by electrostatic interactions. Scaffold candidates were validated for the family *Marseilleviridae* and highlighted a role of H5 as a scaffold protein in poxviruses. Finally, we provide a repertoire of client proteins of the nucleus-like viral factory of mimivirus and demonstrate important sub-compartmentalization of functions including the central dogma. Overall, we reveal a new mechanism for the acquisition of nuclear-like functions entirely based on phase separation and re-classified phylum *Nucleocytoviricota* viral factories as biomolecular condensates.

Keywords

Mimivirus, *Nucleocytoviricota*, Phase Separation, Poxvirus, Viral Factory, Biomolecular condensate, Nucleocytoplasmic large DNA viruses.

Main

Multiple types of viruses generate biomolecular condensates (BMC) which are formed by phase separation (PS) and serve multiple functions in infected cells[1,2]. The viral factories (VFs) of several viruses (DNA and RNA) are formed by PS including members of herpesvirus, adenovirus and the respiratory syncytial virus (reviewed in [1]). Members of the phylum *Nucleocytoviricota* include the *Poxviridae*, the climate-modulating *Emiliania huxleyi* virus and the previously termed Nucleocytoplasmic large DNA viruses (NCLDV)[3-5]. A vast majority of these viruses possess a cytoplasmic exclusive infectious cycle and generate VFs but the molecular mechanism behind their biogenesis is currently unknown[6]. Here, we took advantage of the genetic tools newly developed for GVs to study the nature of the VFs of the phylum *Nucleocytoviricota*[7-10]. Particularly, we focus on the VF of mimivirus since it displays a biphasic nature when imaged by electron microscopy (Figure 1A, Figure S1) and a highly synchronized biogenesis and maturation[6]. We provide evidence that all VFs produced by members of the phylum *Nucleocytoviricota* are generated by PS and predict putative scaffold proteins in all members discovered so far. Moreover, we characterize the VF of mimivirus demonstrating a biphasic behavior accomplished by at least 2 scaffold proteins. Client proteins of the VF of mimivirus were identified and their sub-compartmentalization dissected. Overall, we reclassified VFs of cytoplasmic viruses of the phylum *Nucleocytoviricota* as BMCs and demonstrated that mimivirus VF nuclear-like functions are accomplished by PS.

Nucleocytoviricota viral factories are biomolecular condensates.

While mimivirus VFs successfully seclude DNA and putatively over 300 proteins[11], clear limiting shells are absent when imaged by electron microscopy (Figure 1A). Moreover, previous reports indicated that the mimivirus infection starts with the formation of condensates (originally interpreted as transport vesicles) that coalesce to form the VFs[6]. Coalescent events of the electron-dense inner layer (IL) of the VF can also be detected when superinfected cells are imaged by EM (Figure S1). These observations suggest that a PS phenomenon is involved in mimivirus VFs formation. Thus, to confirm the membrane-less organelle-like nature of these structures, we treated viral factories with 10% 1,6-hexanediol for 10 minutes[2]. Mimivirus VFs partially dissolved during the treatment though smaller factories could be detected occasionally

(Figure 1B-C). Moreover, treatment with 10% 1,6-hexanediol also dissolved the VFs of noumeavirus, pithovirus, pacmanvirus and vaccinia virus (*Poxviridae*) (Figure 1B-D), suggesting a shared mechanism for VF formation across the cytoplasmic viruses in the phylum *Nucleocytoviricota* (Fig 1E). We have previously shown that mollivirus major capsid protein (MCP) accumulates in an uncharacterized sub-compartment in infected cells, before being incorporated into the viral particles[9]. This localization could also be disrupted by 1,6-hexanediol treatment indicating that particle formation at the cytoplasm of infected cells also depends on PS for nuclear viruses (Fig 1C, 1E and 1F).

To gather information associated with the fine ultrastructure of VFs, we fluorescently labelled proteins enriched in a previous proteome of mimivirus VFs[11] and identified two clear sub-compartments (Figure 1G). Moreover, live-cell imaging revealed that upon contact, the outer layer (OL) of two independent VFs fused and never separated during the remaining recording time (Figure 1H). Finally, the number of IL of the VFs linearly increased with the multiplicity of infection (MOI) up to a MOI of approximately 6, strongly suggesting that each genome unit delivered into the cytoplasm of the host (enclosed in the so-called "core") acts as a nucleation point for PS (Figure 1I). Taken together, we concluded the VF of mimivirus (and likely the VFs of all members of the phylum *Nucleocytoviricota*) display characteristics of BMC.

At least two scaffold proteins play key roles in mimivirus viral factory's phase separation.

PS is driven by the multivalency of proteins termed scaffold proteins[12]. Scaffolding proteins are abundant in BMCs and tend to contain intrinsically disordered regions (IDR), which are compacted means to achieve multivalency[12]. In order to identify proteins involved in PS in mimivirus, the over 300 viral proteins present in purified VFs[11] were analyzed for the presence of intrinsically disordered regions (IDR) using flDPnn[13]. Forty-three proteins with IDRs were identified. Two of those proteins (Figure 2A) were enriched in three co-immunoprecipitations (using formaldehyde as a crosslinker agent) of VF proteins (R562, R505 and R336/R337), indicating some degree of proximity with all these client proteins (Supplementary Table 1). Expression in *Acanthamoeba castellanii* of R561 (termed Outer Layer Scaffold 1 (OLS1)) demonstrated it achieved PS in the amoeba cytoplasm (Figure 2B). In contrast, R252 (termed Inner Layer Scaffold 1 (ILS1)) displays a cytoplasmic diffuse localization (Figure 2B). When

cells expressing the two proteins were infected by mimivirus, both proteins re-localized to the viral factories either to the OL (*i.e.* OLS1) or the IL (*i.e.* ILS1) (Figure 2B). Control cells expressing only GFP or RFP did not display such re-localizations (Figure S2A). Importantly, VF client proteins like R336/R337 localized not only to the VF OL but also to BMC formed by the overexpressed OLS1, strongly indicating that this protein is a scaffolding protein forming the OL of the VF (Figure S2B). Moreover, when OLS1 and ILS1 were co-expressed, ILS1 acted as a client protein and was recruited to OLS1 BMC (Figure S2C). Considering that ILS1 binds DNA[14], we reasoned that upon recruitment of ILS1 to the OL of the VF, ILS1 would enter in contact with the viral DNA, where it could achieve PS. To test this hypothesis, we expressed and purified ILS1 in *E. coli* (Figure S2D) and analyzed PS in the presence or absence of DNA (Figure 2C). PS was only observed in the presence of DNA while the protein alone remained soluble (Figure 2C). Moreover, DNA concentrations highly impacted the nature of PS (Figure S2E). Networks were mostly seen at lower DNA concentrations, while large gels appeared at higher concentrations (Figure S2E). ILS1 concentration did not impact the nature of PS (Figure S2F) but rather the speed at which it appeared. Both linear and circular DNA equally triggered PS but mimivirus genomic DNA triggered larger gel formation with lower DNA concentration (Figure S2G). Similar results were obtained with the recombinant mCherry fused ILS1(Figure S2H). RNA did not trigger ILS1 PS (Figure 2C).

In contrast to ILS1, recombinant OLS1 (Figure S2D) made PS independently of other macromolecules (Figure 2D). PS depended on OLS1 concentrations (Figure S3A) and salt concentrations (Figure S3B). Similar to *in vivo* conditions, *in vitro* OLS1 BMC recruited ILS1 as a client protein (Figure S3C). Finally, addition of DNA to the mixture of both proteins triggered a biphasic PS which was achieved independently of the order in which the three components were added to the mixture (Figure 2E). Altogether, OLS1, ILS1 and DNA are enough to trigger VF-like biphasic PS *in vitro*.

In order to attempt gene knockout of both genes, we generated *Acanthamoeba* transgenic lines expressing a codon-optimized version of each protein for trans-complementation[10]. Gene knockout of OLS1 was achieved in trans-complementing cells and clonality of recombinant viruses was demonstrated by genotyping (Figure 3A). A significant reduction of viral particle formation was observed upon deletion of the gene in non-complementing cells (Figure 3B). Moreover, VF formation and/or growth was significantly inhibited as shown by DAPI staining of

infected cells (Figure 3C and 3D). This is confirmed by a lower viral DNA accumulation during infection (Figure 3E). Moreover, depletion of OLS1 also reduced fusion events of VFs upon superinfection as shown by the presence of multiple independent VFs in the same cells (Figure 3C and 3F). However, despite this strong phenotype, this gene is not essential. Thus, either the OL of the VF is composed of multiple scaffold proteins or the OL of the VF is not indispensable for productive mimivirus infection. In order to distinguish between these two hypotheses, we analyzed the mutant infectious cycle by electron microscopy (Figure 3G). VFs of *ols1* KO viruses lack any visible OL, indicating that this compartment of the VF is dispensable for productive infection. Moreover, endogenous tagging of client proteins of the OL of the VF (like R336/R337 or R322) displayed a cytoplasmic localization upon deletion of *ols1* (Figure 3H and S3D-E), strongly suggesting that the role of the OL of the VF of mimivirus is to concentrate components important for VF functions at the IL or its periphery. The observed phenotypes in wild-type cells infected by *ols1* KO mutant also support the lack of additional scaffold proteins to replace OLS1.

On the other hand, we were unable to obtain clonal *ils1* KO viruses. Using the trans-complementing line, we demonstrated that *ils1* is an essential gene (Figure 3I). Future efforts will be directed at optimizing conditional depletion systems in GVs to study the function of essential genes by reverse genetics.

A conserved molecular grammar allows the identification of viral factory scaffold proteins throughout the phylum *Nucleocytoviricota*.

The molecular grammar of an IDR refers to compositional bias and sequence patterns in their primary structure[15,16]. This grammar allows condensation and specific recruitment of molecules, including client proteins[17-22]. We thus reasoned that due to the plethora of client proteins recruited to the VFs, despite major changes in the primary sequence of their scaffold proteins, the molecular grammar of the condensate would not easily change. Concordantly, host-expressed mimivirus (Imitervirales) OLS1-GFP and ILS1-RFP re-localized to noumeavirus (Pimascovirales) VF upon infection (Figure 4A). Thus, in order to identify scaffold proteins in all *Nucleocytoviricota*, we first predicted the "IDRome" encoded in representative genomes of isolated viruses from each viral order and extended the analysis to metagenomes from the giant virus database[23], the permafrost[24] and to the recently discovered Egovirales[25] (Supplementary

Table 2). We then tested existing tools to classify such IDRs. Since the IL of the VF is only present in members of the *Mimiviridae*, we excluded ILS1 from the following analysis and focused bioinformatic computations on OLS1. While tools to predict phase separation exist (ParSe v2[26] and Molphase[27]), they predicted many candidates with the potential of achieving PS (Figure S4A). Even using a stringent threshold (0.9), molphase predicts 69% of mimivirus proteins with an IDR to do phase separation and ParSe v2, 31% of mimivirus disordered proteins. Thus, we customized specific methods developed to study the biochemistry of nucleolar phase separation relying on Nardini and CIDER[22] to increase the prediction specificity. The first IDR of OLS1 and its detectable homologues were used to determine important features separating these IDRs from the rest of the IDRome of mimivirus. Compared to the previously published method[22], the number of scrambles for Nardini Z-score calculations was reduced (Figure S5A) and block sizes of certain amino acids were slightly adapted to allow better discrimination of OLS1 (Figure S5B). Using those 98 features in total (36 from Nardini and 62 from CIDER), we created machine learning classifiers that identified 4 candidate scaffold proteins in noumeavirus (Figure S4A). To experimentally challenge these predictions, the four genes were codon-optimized and expressed in the amoeba. Two of them showed a localization coherent with BMCs, which could be disrupted using 1,6-hexanediol *in vivo* (Figure 4B) and re-localized to the noumeavirus VF upon infection (Figure 4B). Moreover, both proteins also re-localized to the mimivirus VFs upon infection, confirming a shared molecular grammar for PS for these 2 viruses VFs (Figure 4B). In contrast, the other two protein candidates did not spontaneously form BMCs in the amoeba cytoplasm (Figure S4B).

We then utilized the newly confirmed noumeavirus scaffold proteins (NMV_095 and NMV_238) and detected homologs to re-train the predictive machine learning classifier. To select the features for the classifier, we compared the features values of the three scaffold proteins IDRs and homologs to the rest of the combined IDRome of mimivirus and noumeavirus (Figure S5C-D). Final predictions of scaffold proteins in all orders of *Nucleocytoviricota* were then generated with this optimized classifier (Supplementary Table 2). Clear segregation of resulting candidate IDRs can be observed on the unsupervised UMAP representation of the features previously selected during the classifier training (Figure 4C) and putative scaffold proteins could be identified to be encoded by members of all *Nucleocytoviricota* orders (Figure 4C-D). More specifically, in all reference genomes from cytoplasmic viruses, we identified one to five

candidates except for African swine fever virus where no candidates could be identified (Figure 4D). In all these (cytoplasmic) orders, between 4.4% (*Algalvirales*) to 5.4% (*Chitovirales*) of all predicted IDRs were classified as candidate scaffold proteins for PS and VFs generation. The only exception was the *Egovirales*, scoring 8.7% of positive proteins in their respective IDRomes. In *Pimascovirales*, homologous proteins were identified as scaffolds in pithovirus sibericum and cedratvirus kamchatka (pv_12 and ck125, 39% identity/63% similarity), highlighting the consistency of the method. Similarly, orthopoxvirus' H5 protein was identified as the only candidate both in vaccinia virus and monkeypox virus (93% identity). On the other hand, in *Pandoravirales* (which present a nuclear DNA replication[28] and apparently lack VFs), around 10 candidate proteins were identified. Importantly, all candidate VF scaffold IDRs predicted in all *Nucleocytoviricota* are distributed on the UMAP without order segregation, further reinforcing the shared molecular grammar for the cytoplasmic VFs (Figure 4C). As expected, scaffolds IDRs have a low Anchor2 score meaning they are predicted to stay disordered. Meanwhile, a significant portion of other IDRs have a higher Anchor2 score indicating that they likely gain conformational order upon binding to a molecule (Figure S6A).

The most discriminative features in our classification and thus, the features associated with the molecular grammar of the VFs, were related to charge segregation (Figure 4E-G). OLS1, NMV_095 and NMV_238 display large patches of positively and negatively charged residues (Figure S7) resulting in high Nardini positive-positive and positive-negative Z-scores. The kappa parameter[29] is also higher for the confirmed scaffold proteins and predicted candidates (Supplementary Table 2), further pointing towards a high segregation of charges. Regardless, kappa is not part of the final classifier as the parameter is correlated to the positive-negative Nardini Z-score in the scaffold protein training set (Figure 4E). Within the negative charges, glutamate seems more important as confirmed scaffold IDRs and candidates have 15% of this residue (+/- 6) while the rest of the IDRs have 6% (+/- 8). This feature is also very important to discriminate OLS1 from ILS1 (which has only 1.4% of glutamate (and other negative residues)), probably reflecting its inability to perform PS without DNA (Figure 2C). There is another group of IDRs at the center-bottom of the UMAP that scores high in positive segregation (Figure 4G). This group resembles the candidate scaffolds in some classifier metrics but presents a low fraction of E residues for instance (Figure S6B). Like for scaffold proteins of the nucleolus, Blocks of K are higher in VF scaffold candidates than in other IDRs (0.7 +/- 0.11 vs 0) but this

feature was not included in the classifier due to its high variability in scaffolds and homologs. In addition, scaffold IDRs and candidates have a relatively higher E/D ratio (0.25 +/- 0.26 vs 0 +/- 0.45) and a higher K/R ratio (0.48 +/- 0.36 vs 0.07 +/- 0.53).

Although the two first metrics are the most discriminant (Figure 4F-G), the other features are essential as they help to reduce the number of candidates from 6539 if we create a classifier based only on the two first metrics, to 2967 candidates with the final classifier considering those 11 metrics (Supplementary Table 2). It is the combination of all those features (Figure 4F-G, Figure S6B) that makes a specific discrimination possible.

Identification of client proteins demonstrates sub-compartmentalization of functions.

While a previous study reported proteins tentatively localized at the VFs of mimivirus[11], it did not differentiate proteins associated with each sub-compartments of the VFs and high rates of false positives were obtained when localization of some of these proteins was assessed by endogenous tagging (Figure S8A). In order to fill this gap, we performed immunoprecipitation (IP) of the two scaffold proteins and identified co-purified proteins by mass spectrometry (MS)-based proteomics (Figure 5A and Supplementary Table 3). Several proteins found enriched with OLS1 and/or ILS1 were endogenously tagged to confirm their localization (Figure 5B and Figure S8B). Importantly, none of the false positive examples identified from the previous proteome study were detected in these co-IPs (Supplementary Table 3). Moreover, an enrichment of proteins localized to the OL of the VF were co-purified with OLS1, while IL proteins or virion proteins were co-immunoprecipitated majorly with ILS1 (Figure 5A). Importantly, while this study shows a low rate of false positive identifications, other proteins not detected in the immunoprecipitations are still localized at the VF (Figure 5C). Thus, future efforts will be needed to identify the full proteome of the VFs of mimivirus.

Several host ribosomal proteins were enriched in the co-IP with OLS1-GFP (but not with ILS1-RFP), suggesting some proximity between the OL of the VF and the ribosomes (Figure 5A and Supplementary Table 3). Regardless, VFs are thought to segregate replication and transcription from translation[11]. In order to test the current consensus, we generated recombinant viruses or amoebas encoding tagged versions of several proteins involved in the central dogma (Figure 6A-D and Figure S8A). As previously suggested[11], host ribosomes (Figure 6C and Figure S9A) and

the viral translation-associated protein SUI1 (Figure 6D) were excluded from the VFs. On the other hand, the virally encoded eIF4e localized both at the cytoplasm of the host and the OL of the VF (Figure 6D). In eukaryotes, besides its classical cytoplasmic function in translation initiation, eIF4E also localizes at the nucleus where it participates in the export of a subset of mRNA[30]. If such export function is conserved in the virally encoded eIF4E remains to be explored, but it would explain the dual localization of the protein. Surprisingly, while proteins associated to replication and transcription are incorporated in the VFs, replication proteins localized to the OL of the VF while transcription proteins accumulated at the IL (Figure 5A and Figure 6A-B). Taken together, these data indicate important sub-compartmentalization of the processes associated to the central dogma. Pulse-labeling of mRNAs using EU for 15 minutes strongly suggests that mRNA production site locates at the IL of the VF (Figure 6E), colocalizing with the RNA polymerase (Figure 6B). On the other hand, we were unable to efficiently label DNA by EdU in wild type viruses (Figure S9B). We reasoned that due to the high AT richness of mimivirus genome[31], *de novo* synthesis of dTMP would be particularly efficient in infected cells, allowing thymidine to outcompete EdU for its incorporation into the DNA (Figure S9C). Thus, we generated knockout recombinant viruses on the virally encoded thymidylate synthase (TS), concordantly blocking viral *de novo* synthesis of dTMP (Figure S9D-E). EdU labeling significantly improved in these viruses allowing to visualize DNA replication with pulses of EdU labelling as short as 5 minutes (Figure 6F and Figure S9F). Similarly to what was observed in vaccinia virus[32], pulse labeling of DNA replication in late VF resulted in lower intensity of fluorescence than labelling in early VF (Figure 6F-G). These data strongly indicate that DNA replication decreases during late stages of infection. Moreover, 5-minute labelling with EdU showed an enrichment of newly synthesized DNA at the periphery of the OL of the VF (Figure 6F). Concordantly, DNA replication likely occurs at the interface between the IL and the OL where DNA and the DNA polymerase get in contact. Finally, mRNA processing (including capping and Poly(A) synthesis) localized at the IL of the VF, indicating that maturation of pre-mRNA occurs at the same sub-compartment as transcription (Figure 6H). Overall, a significant sub-compartmentalization of functions is observed at mimivirus VFs.

Discussion

The molecular grammar of an IDR of a scaffold protein determines the nature of the interaction to achieve PS and the selective recruitment of client proteins[17-22,33]. Here, we demonstrate that members of the *Nucleocytoviricota* phylum share a common molecular grammar which allows the formation of their VFs. Thus, the common ancestor of all *Nucleocytoviricota* already possessed a VF with this molecular grammar. Such conservation is likely driven by the complex number of client proteins recruited to the VFs, which would need to change simultaneously their biochemical/biophysical properties if the molecular grammar of the VF suddenly changes. Such a scenario is parsimoniously unlikely. Regardless, it is unclear if all scaffold protein IDRs share a common origin and diversification occurred by shuffling protein fragments[34] or if the same molecular grammar emerged in multiple IDRs on different occasions by convergent evolution[35]. Which advantages the virus gains by modifying the scaffold proteins that form their VFs remains to be addressed, but might be associated with emergent traits of different VFs (including the appearance of an IL in mimivirus or the endoplasmic reticulum (ER) wrapping in poxviruses[36]). Importantly, such grammar allowed us to identify previously neglected but well-characterized scaffold proteins in other members of the phylum. In *Poxviridae*, H5 was predicted as the sole candidate to be a scaffold protein for PS. H5 is an essential protein for the infectious cycle of vaccinia virus[37] and was coined as a hub protein due to its importance in DNA replication, transcription and virion morphogenesis[38]. All those phenotypes correlate with a scaffolding function for PS. Moreover, H5 binds DNA[38] (function which is modulated by phosphorylation[39]) and localizes as puncta upon heterologous expression in mammalian cells (indicating spontaneous PS of the protein)[37]. Interestingly, when vaccinia virus uncoating occurs, early viral proteins associated with DNA replication localize to cytoplasmic puncta, including H5[40,41]. We propose that these puncta (known as prereplication foci) are likely formed by PS using H5 as a scaffold protein. Upon maturation of the prereplication foci, VF gets surrounded by the ER membranes[36]. Regardless, H5 strongly accumulates inside the VF[41,42] and the VFs dissolve in the presence of 1,6-hexanediol. This supports the idea that PS is a major driver of the VF formation regardless of the ER wrapping. Moreover, it has been proposed that H5 would be a key component for VF enlargement and wrapping by the ER[40]. Overall, previously published data on H5 strongly supports its role as an unrecognized scaffold protein for PS. In *Marseilleviridae*, at least two scaffold proteins localize to the VFs with differential transcriptional expression patterns[43]. This allows us to hypothesize that NMV_095 might initiate VF formation while

NMV_238 would allow its expansion and maturation. Further experiments will be needed to corroborate this hypothesis. In *Pandoravirales*, multiple IDRs containing a similar molecular grammar than VF scaffold proteins were identified. Regardless, these viruses do not form VFs and transfer their DNA into the nucleus of their host. In evolutionary terms, it is parsimonious to assume that a fully cytoplasmic infectious cycle style of the majority of the members of the phylum originated prior to the nuclear one[44] since the origin of *Nucleocytoviricota* predated the origin of eukaryotes (and thus, the nucleus)[45]. Thus, during the transition from cytoplasmic to nuclear viruses, the transfer of the DNA into the nucleus would generate a VF which is no longer needed to achieve replication and transcription but would still keep the functions associated to virion morphogenesis. Regardless, further experiments would be needed to characterize such "Virion Factories" potentially assembled by nuclear GVs.

Mimivirus VFs are highly compartmentalized organelle-like structures. Biogenesis of the VFs starts by utilizing the cores as nucleating points. Mimivirus VFs then develop into multilayered structures that contain at least two distinctive phases. The OL of the VF, formed by OLS1, acts as a selective barrier and recruits VF proteins. Importantly, while we hypothesize that the OL of mimivirus VFs is the phylogenetically conserved phase between different *Nucleocytoviricota*, OLS1 is dispensable in mimivirus. We theorize that the presence of the IL allows to protect the genome of the virus in absence of the OL and, despite losing the ability of selectively recruiting proteins to the VF and considering the permissive conditions of the laboratory, the IL is sufficient to achieve successful infection. The DNA replication machinery localizes to the OL of the VF and maximizes DNA replication only at the interphase between OL and IL. This interphase is also the site of assembly of virions and the competition between these two processes might at least partially explain the decrease in DNA synthesis during late stages of the infection cycle. The IL of the VF contains proteins associated to transcription which are also packaged into the virions to establish a new cycle of infection (including the RNA polymerase, transcription factor, RNA processing, etc.). Thus, the IL seems to be a compartment analogous to the internal content of the virion core, which is sufficient to re-start RNA transcription of early genes upon infection of a new cell[46]. Moreover, ILS1 has recently been proposed to work on mimivirus DNA condensation for its incorporation in the viral particle[14]. Finally, translation occurs outside of the VF, a feature differentiating mimivirus from vaccinia virus[47]. Importantly, how mRNAs are exported from the VFs is still unknown. Regardless, since neither ILS1 nor OLS1 require RNA for PS, a simple

model where RNA is not retained by the IL or OL of the VF can be envisioned. In such case, diffusion would be sufficient to deliver mRNAs into the cytoplasm of the infected cell.

Overall, *Nucleocytoviricota* VFs are BMCs with a shared phylogenetic history and a common molecular grammar. Such discovery raises new questions including how cytoplasmic viruses interact upon infection of the same host cell (can VFs from different viruses fuse?), how does BMCs accommodate spatiotemporally the different functions required for VFs multiple roles and open the door to the development of generalist drugs to inhibit *Nucleocytiviricota* viral infections.

Author contributions

S.R., conceptualization, methodology, validation, formal analysis, investigation, data curation, writing (original draft, review and editing), software, visualization; A.S., conceptualization, methodology, validation, formal analysis, investigation, data curation, writing (original draft, review and editing), software, visualization; A.L., investigation and data analysis; L.D., investigation and data analysis; C.G., investigation and data analysis; F.T., investigation and data analysis; L.B., investigation and data analysis; N.O-D., investigation, data analysis and writing (review & editing); Y.C., supervision, methodology, validation, formal analysis and writing (review & editing); M.B., investigation, data analysis, supervision and writing (review & editing); M.L., supervision and writing (review & editing); S.J., supervision and writing (review & editing); C.A., conceptualization, supervision, writing (review & editing) and funding acquisition; H.B., conceptualization, methodology, validation, formal analysis, investigation, data curation, visualization, writing (original draft, review and editing), supervision, project administration and funding acquisition.

Acknowledgments

The authors thank Matias Estaras Hermosel for critical discussion to improve the manuscript. We thank the members of the PiCSL-FBI core facility (Nicolas Brouilly, Fabrice Richard and Aïcha Aouane), IBDM, AMU-Marseille; and on the IMM imaging platform (Artemis Kosta and Hugo Le Guenno). The proteomic experiments were partially supported by Agence Nationale de la Recherche under projects ProFI (Proteomics French Infrastructure, ANR-10-INBS-08) and GRAL, a program from the Chemistry Biology Health (CBH) Graduate School of University Grenoble Alpes (ANR-17-EURE-0003). NO y MB gratefully acknowledge the Advanced

Bioimaging Unit at the Institut Pasteur Montevideo & Universidad de la República for their support and assistance in the present work.

Declaration of interests

The authors declare no competing interests.

Methods

A. castellanii growth and virus production

The following viral strains have been used in this study: Acanthamoeba polyphaga mimivirus[31], noumeavirus[44], pithovirus sibericum[48], pacmanvirus lost city (*manuscript in preparation*), mollivirus sibericum[49]. Ten infected 75 cm² tissue-culture flasks plated with fresh *Acanthamoeba cells* were used for virus production. After lysis completion, the cultures were recovered, centrifuged 5 min at 500 × g to remove the cellular debris, and the virus was pelleted by a 45 min centrifugation at 6,800 × g prior purification. The viral pellet was then resuspended and washed twice in PBS and layered on a discontinuous CsCl gradient (1.2/1.3/1.4/1.5 g/cm3), and centrifuged at 100,000 × g overnight. An extended protocol is shown in [50].

Vaccinia virus.

The Modified Vaccinia virus Ankara (MVA) strain was propagated in BHK-21 cells. Viral stock was generated in 25 cm² tissue-culture flasks with cell monolayers at a multiplicity of infection (MOI) of 0.1. After a 1-hour adsorption period, the inoculum was removed, and the cells were incubated at 37°C in a humidified atmosphere with 5% CO□. The cultures were collected when 80-90% of the cell monolayer exhibited cytopathic effects (CPE), typically 2-3 days post-infection. The cultures were centrifuged at 1500 xg for 5 minutes, after which the supernatant was aliquoted and stored at -80°C. The viral titer was determined using a TCID50 endpoint dilution assay.

For experiments involving 1,6-hexanediol treatment, viral infections were carried out in Vero cells using an MOI of 10.

Both BHK-21 and Vero cells were cultured in Dulbecco's Modified Eagle's Medium (DMEM) supplemented with 10% fetal bovine serum (FBS) and an antibiotic-antimycotic solution (penicillin 100 units/mL, streptomycin 100 µg/mL, and amphotericin B 0.25 µg/mL). During viral infections, the FBS concentration was adjusted to 1%.

Acanthamoeba castellanii (Douglas) Neff (American Type Culture Collection 30010TM) cells were cultured at 32 °C in 2% (wt/vol) proteose peptone, 0.1% yeast extract, 100µM glucose, 4mM $MgSO_4$, 0.4mM $CaCl_2$, 50 µM $Fe(NH_4)_2(SO_4)_2$, 2.5 mM Na_2HPO_4, 2.5 mM KH_2PO_4, pH 6.5 (home-made PPYG) medium supplemented with antibiotics [ampicilline 100 µg/mL, and Kanamycin 25 µg/mL]. 100 µg/mL Geneticin G418 or Nourseothricin was added when necessary.

Generation of DNA constructs

Vectors for endogenous tagging.

vAS1 plasmid was utilized for endogenous tagging[7]. 500 bp homology arms were introduced at the 5' and 3' end of the selection cassette in order to induce homologous recombination with the viral DNA. Each cloning step was performed using the Phusion Taq polymerase (ThermoFisher) and InFusion (Takara). Prior to transfection, plasmids were digested with ApaI/EcoRI/HindIII and NotI. Primers utilized are shown in Supplementary Table 4.

Second copy vectors for expression in A. castellanii.

R561, R252, NMV_95, NMV_141, NMV_227 and NMV_238 encoding genes were codon optimized for amoeba expression and amplified by PCR to be cloned into different amoeba expression vectors (PAM1, PAM2, PAM3, PAM10 or Vc241[7]). The plasmid was linearized by NdeI and the gene was inserted using InFusion Takara. Primers utilized are shown in Supplementary Table 4.

Vectors for gene knockout of ols1 (r561) and ils1 (r252).

vHB47 was used as the plasmid for gene knock-out[7,8]. 500 bp homology arms were introduced at the 5' and 3' end of the selection cassette to induce homologous recombination with the viral DNA. Each cloning step was performed using the Phusion Taq polymerase (ThermoFisher) and InFusion (Takara). Before transfection, plasmids were digested with ApaI/EcoRI/HindIII and NotI. Primers utilized are shown in Supplementary Table 4.

Vector for gene knockout of thymidylate synthase.

vAS1 was used as the plasmid for gene knock-out[7,8]. 500 bp homology arms were introduced at the 5' and 3' end of the selection cassette to induce homologous recombination with the viral

DNA. Each cloning step was performed using the Phusion Taq polymerase (ThermoFisher) and InFusion (Takara). Before transfection, plasmids were digested with HindIII and NotI. Primers utilized are shown in Supplementary Table 4.

Establishment of viral lines

Generation of recombinant viruses

Recombinant viruses were generated as described step by step in [7]. Briefly, $1.5x10^5$ *Acanthamoeba castellanii* cells were transfected with 6 µg of linearized plasmid using Polyfect (QIAGEN) in phosphate saline buffer (PBS). One hour after transfection, PBS was replaced with PPYG and cells were infected with mimivirus for 1 hour with sequential washes to remove extracellular virions. 24h after infection the new generation of viruses (P0) was collected and used to infect new cells. An aliquot of P0 viruses was utilized for genotyping in order to confirm integration of the selection cassette. New infection was allowed to proceed for 1 hour, then washed to remove extracellular virions and nourseothricin and/or geneticin was added to the media. Viral growth was allowed to proceed for 24 hours. This procedure was repeated one more time before removing the nourseothricin and/or geneticin selection to allow recombinant viruses to expand more rapidly. Once, viral infection was visible, selection procedure was repeated one more time. Viruses produced after this new round of selection were used for genotyping and cloning. Selection utilized for each virus generation is indicated in the "Generation of DNA constructs" section.

Cloning and genotyping. Cloning and genotyping of recombinant viruses are extensively described in [7]. Briefly, 50,000 *A. castellanii* cells were seeded on 6 well plates with 2 mL of PPYG. After adhesion, viruses were added to the well at a multiplicity of infection (MOI) of 2. One-hour post-infection, the well was washed 5 times with 1mL of PPYG and cells were recovered by well scraping. Amoebas were then diluted until obtaining a suspension of 1 amoeba/µL. 1 µL of such suspension was added in each well of a 96-well plate containing 1,000 uninfected *A. castellanii* cells and 200 µL of PPYG. Wells were later monitored for cell death and 100 µL collected for genotyping. Genotyping was performed using Terra PCR Direct Polymerase Mix (Takara) following manufacturer specifications. Primers utilized for each genotyping are detailed in Supplementary Table 4.

Protein expression and purification

Protein expression and purification were performed as previously described[14]. Briefly, cultures were grown in Luria-Bertani (LB) medium containing ampicillin until an optical density of 0.5-0.6 (600nm). Bacterial expression was then induced with 0.3 mM isopropyl-1-thio-β-d-galactopyranoside (IPTG) and bacteria were incubated at 16°C for 15 h with constant shaking. Cells were then collected and centrifuged at 4,000 g prior to resuspension in 50 mM Tris pH 7.4, 5 mM imidazole, 1 M NaCl, 5% glycerol, 1 mM Phenylmethylsulfonyl fluoride (PMSF) and 1 mM benzamidine hydrochloride. Cells were lysed by sonication prior to clearing by centrifugation at 16,000g for 30 min. Soluble fraction was loaded on a Hi-Trap Chelating HP 1 ml pre-packed column (GE Healthcare/Cytiva). Elution was carried out in 50 mM Tris pH 7.4, 5 mM imidazole, 500mM NaCl, 5% glycerol, 1 mM Phenylmethylsulfonyl fluoride (PMSF), 1 mM benzamidine hydrochloride. Protein fractions obtained were pooled and later dialyzed with 20 mM Tris pH 7.5, 100 mM NaCl and 5% glycerol. Purified proteins were concentrated using Amicon® Ultra 15 ml centrifugal filters and stored in aliquots at -80°C. Protein concentrations were determined using the nanodrop and their theoretical epsilon.

Phase separation assays

All droplet formation assays were performed in absence of crowding agents. Proteins were diluted into specified buffers in a final assay volume of 100 µL. When indicated RNA or DNA was added to the mixture. Samples were visualized on a 96 well non-binding microplates (Greiner bio-one).

Immunofluorescence and fluorescent microscopy

A. castellanii cells were grown on poly-L-lysine coated coverslips in a 12-well plate, infected or not with viruses and fixed with PBS containing 3.7% formaldehyde for 20 min at room temperature. When required, immunofluorescence was performed as described step by step in [7]. After three washes with PBS buffer, coverslips were mounted on a glass slide with 4 µl of VECTASHIELD mounting medium with DAPI and the fluorescence was observed using a Zeiss Axio Observer Z1 inverted microscope using a 63x objective lens associated with a 1.6x Optovar for DIC, mRFP or GFP fluorescence recording.

Vero cells were grown on poly-L-lysine treated 96-well plate, infected or not with viruses and fixed with 4% formaldehyde in PBS (Fixative Solution, Invitrogen) with DAPI for 15 min at room temperature. The fluorescence was observed using a 40x objective lens in a Olympus IX81 inverted microscope using the µManager software.

1,6-hexanediol treatment

To disrupt viral factories, 1,6-hexanediol (240117, Sigma-Aldrich) was diluted in cell culture media at 10% w/v as previously described[2]. Cell culture media were replaced with media containing 10% 1,6-hexanediol or fresh culture media and incubated for 10 min at 32 or 37 °C before fixation.

Virion production quantification

Optical density was utilized for viral quantification as previously described[7]. Purity of the viral samples were analyzed by microscopy[50] or genotyping[7] as previously described.

DNA replication

Viral genomes or gDNA from infected amoebas were purified using Wizard genomic DNA purification kit (PROMEGA). To determine the amplification kinetic, the fluorescence of the EvaGreen dye incorporated into the PCR product was measured at the end of each cycle using SoFast EvaGreen Supermix 2× kit (Bio-Rad, France). A standard curve using gDNA of purified viruses was performed in parallel for each experiment. For each point, a technical triplicate was performed. Primers utilized are shown in Supplementary Table 4.

Electron microscopy imaging

Extracellular virions or *A. castellanii*-infected cell cultures were fixed by adding an equal volume of PBS with 2% glutaraldehyde and 20 min incubation at room temperature. Cells were recovered and pelleted 20 min at 5,000 × g. The pellet was resuspended in 1 mL PBS with 1% glutaraldehyde, incubated at least 1 h at 4 °C, and washed twice in PBS prior coating in agarose and embedding in Epon resin. Each pellet was mixed with 2% low melting agarose and centrifuged to obtain small flanges of approximately 1mm^3 containing the sample coated with agarose. These samples were then prepared using the osmium-thiocarbohydrazide-osmium method: 1 h fixation in 2% osmium tetroxide with 1.5% potassium ferrocyanide, 20 min in 1%

thiocarbohydrazide, 30 min in 2% osmium tetroxide, overnight incubation in 1% uranyl acetate, 30 min in lead aspartate, dehydration in increasing ethanol concentrations (50, 70, 90 and 100% ethanol) and embedding in Epon-812. Ultrathin sections of 70 nm were observed using a FEI Tecnai G2 operating at 200 kV[44].

Immunoprecipitation

6 hours post-infection *Acanthamoeba castellanii* infected cells were harvested, washed in PBS and lysed in co-immunoprecipitation buffer (0.2% v/v Triton X-100, 50 mM Tris-HCl, pH 8, 150 mM NaCl) in the presence of a protease inhibitor cocktail (Roche). Cells were sonicated on ice and centrifuged at 14,000 r.p.m. for 30 min at 4°C. Supernatants were then subjected to immunoprecipitation using anti-HA, anti-GFP, or anti-RFP antibodies, as previously described[51].

MS-based proteomic analyses

Proteins eluted from co-IP experiments were either separated by SDS-PAGE (HA-tagged proteins and WT control, one replicate per condition) or stacked (GFP- and RFP-tagged proteins and respective controls, three replicates per condition) in the top of a 4-12% NuPAGE gel (Invitrogen) before Coomassie blue staining and in-gel digestion using modified trypsin (Promega, sequencing grade) as previously described[52]. For co-IP experiments with HA-tagged proteins and WT control, the bands corresponding to large chains of immunoglobulins were prepared and analysed separately from the rest of the samples. The resulting peptides were analyzed by online nanoliquid chromatography coupled to MS/MS (Ultimate 3000 RSLCnano and Q-Exactive HF, Thermo Fisher Scientific) using gradients of 140 min, 35 min or 80 min for the eluates of HA-tagged proteins and of the WT control, the bands corresponding to the immunoglobulin large chains in HA-co-IPs, and GFP- and RFP-tagged proteins and the corresponding controls, respectively. The MS and MS/MS data were acquired using Xcalibur (Thermo Fisher Scientific). The mass spectrometry proteomics data have been deposited to the ProteomeXchange Consortium via the PRIDE[53] partner repository with the dataset identifier PXD054803.

Peptides and proteins were identified by Mascot (version 2.8.3, Matrix Science) through concomitant searches against the following homemade databases: *A. castellanii* nuclear genome (17’625 sequences), *A. castellanii* mitochondrial genome (40 sequences), mimivirus (979 sequences), and contaminants classically found in proteomic analyses (keratins, trypsin… 250

sequences). Trypsin/P was chosen as the enzyme and two missed cleavages were allowed. Precursor and fragment mass error tolerances were set at respectively at 10 and 20 ppm. Peptide modifications allowed during the search were: Carbamidomethyl (C, fixed), Acetyl (Protein N-term, variable) and Oxidation (M, variable). The Proline software[54] (version 2.3) was used for the compilation, grouping and filtering of the results (conservation of rank 1 peptides, peptide length ≥ 6 amino acids, false discovery rate of peptide-spectrum-match identifications < 1%[55], and minimum of one specific peptide per identified protein group). Proline was then used to perform a MS1-based label-free quantification of the identified protein groups based on specific and razor peptides.

Statistical analysis was performed using the ProStaR software[56]. Proteins identified in the contaminant database were discarded. For the dataset of HA-tagged proteins, after log2 transformation, abundance values were normalized using median centering. To be considered enriched with a HA-tagged bait protein, a protein must show a normalized abundance at least four times higher in the bait protein eluate than in the WT control eluate and be identified with a minimum of three spectral counts. For the dataset of GFP- and RFP-tagged proteins, proteins detected in less than three replicates of one condition were discarded. After log2 transformation, abundance values were normalized using the variance stabilizing normalization (vsn) method, before missing value imputation (SLSA algorithm for partially observed values in the condition and DetQuantile algorithm for totally absent values in the condition). Statistical testing was conducted with limma, whereby differentially expressed proteins were selected using a log2 (Fold Change) cut-off of 1.6 and a p-value cut-off of 0.01, allowing to reach false discovery rates inferior to 2% according to the Benjamini-Hochberg estimator. Proteins detected in fewer than three replicates in the condition in which they were most abundant were manually invalidated (p value = 1).

EU and EdU labelling

Acanthamoeba cells were grown on glass coverslips and infected with mimivirus at MOI of 10. Incorporation and visualization of EU or EdU was performed as previously described[46] utilizing Click-iT™ EdU Cell Proliferation Kit for Imaging, Alexa Fluor™ 488 dye and Click-iT™ RNA Alexa Fluor™ 488 Imaging Kit Invitrogen. Briefly, after labeling for the specified time with 100 µM EdU or 1 mM EU, cells were fixed with 3.7% paraformaldehyde for 20 min, washed,

permeabilized with 0.5% Triton X-100, and incubated with the Click-iT™ reaction mixture as indicated by the manufacturer.

Statistics and reproducibility

All data are presented as the mean ± s.d. of 3 independent biological replicates (n = 3), unless otherwise stated in the figure. All data analyses were carried out using Graphpad Prism. The null hypothesis (α = 0.05) was tested using unpaired two-tailed Student's t-tests.

Database constitution with molecular grammar of IDRs

A database was constituted with genomes of isolated viruses: African swine fever virus (GCA 000858485.1), cedratvirus kamchatka (GCA 031200085.1), Invertebrate iridescent virus 6 (IIV-6, GCA 000838105.1), marseillevirus (GCF 001806195.1), mimivirus (GCA 000888735.1), mollivirus sibericum (GCF 001292995.1), monkeypox virus Zaire (MPV-ZAI, GCA 000857045.1), noumeavirus (GCF 002005685.1), pandoravirus neocaledonia (GCF 003233915.1), paramecium bursaria chlorella virus 1 (PBCV-1, GCA 000847045.1), pithovirus sibericum (GCA 000916835.1), powai lake megavirus (GCA 002924545.1), vaccinia virus WR (VACCW, GCA 900236015.1). Predicted proteins from the Giant virus database[23] from the 8 large genomes from permafrost metagenomics (PRJEB47746), and from *Egovirales*[25] were also included. Intrinsically disordered regions in all proteins were predicted using MobiDB-lite v3.10.0[57].

The python package Nardini v1.1.1 was used to infer Z-scores for all IDRs based on the positive, negative, polar, hydrophobic, aromatic residues and alanines, prolines or glycines[16]. The Z-scores were inferred from 50,000 scrambles, meaning the sequences are shuffled 50,000 times in order to calculate a Z-score of the real value. The optimal number of scrambles was chosen by comparing Z-scores from 10 to 500,000 scrambles to the Z-scores obtained with 1,000,000 scrambles using the IDRs of mimivirus and homologs of its scaffold proteins.

Compositional data and physical and chemical properties of IDRs were predicted with localCIDER v0.1.21[58] as in [22] with the addition of the kappa estimation. The block lengths of certain residues were also counted similarly to King et al.[22], only counting the residues of interest (no mismatch) and subtracting opposite residues (positives vs negatives, hydrophobic vs polar).

The resulting Nardini and CIDER features were then normalized by subtracting the median and dividing by the inter-quantile range previously calculated on all IDRs, including from metagenomes. All Z-scores and normalized features were then saturated by the sigmoid function[59].

Constitution of the reference IDR and homologs database

Homologs of OLS1, NMV_095 and NMV_238 were recovered in two steps. First, the proteins were aligned to the Giant virus database and the permafrost metagenomes[24] by MMseqs2 v.12[60] with an e-value cutoff of 1e-5. Secondly, sequences were aligned with t-coffee v13.41.0[61] and HMM models were constructed with HMMER v3.3.2[62] and searched for in the metagenomic database. Sequences were considered as homologs only for alignments with e-values $<1e^{-10}$. The homologous proteins were then aligned again with t-coffee. Only IDRs aligned to the reference (OLS1__1, NMV_095__1 or NMV_238__1) with at least 10 overlapping amino acids were kept. Important features were determined in the same way as for the final classifier: wilcoxon signed-rank test from the scipy package v. 1.13.1 were performed to compare reference IDRs and homologs to the rest of the IDRs of mimivirus or noumeavirus. The p-values were corrected by the false_discovery_control function. Only features with a significantly lower variance, given by variance comparison and a levene test, were considered to further ensure that we compared a homogeneous population. For all the significantly relevant features, the Euclidean distance was calculated and IDRs with a distance above a manually set threshold were discarded after inspection of heatmaps presenting those features for each IDR. Three IDRs were then manually removed from the homologs of NMV_238 as they were too distant from the scaffolds in the UMAP. The final homologous IDRs used for the training set are given in Table S7.

Prediction of the VF scaffold proteins

Several classifiers and distance-based methods were created to first, identify candidate proteins in noumeavirus that could correspond to OLS1 and second, predict OLS1 proteins in other genomes with the generalization knowledge gained from noumeavirus candidates' validation. For the exploratory phase, a preliminary distance-based classifier was defined: the Euclidean distance of all normalized IDR features with a q-value under 0.05 to OLS1 was calculated. IDRs with distances smaller than the furthest homolog were considered as candidates. Prediction refinement

was achieved by setting up SVM-based classification: we sorted features on their q-value and created classifiers based on 2 to 20 features not to exceed approximately 1/10th the number of IDRs in the training set to avoid overfitting (Fig.S5D). The Spearman correlation between features in the scaffold proteins (reference IDR and homologs) was assessed and were only considered the most discriminant features above a certain correlation threshold (Figure S5C). We tested SVM classifiers with both linear and rbf kernels, with a C parameter of 5 and a gamma parameter of 5. Data points were weighted according to this scheme: each negative point was given a weight of 20 divided by the number of negative points in the genome, each positive point, a score of 3.33 divided by the number of homologs and the experimentally confirmed scaffold IDRs were given a weight of 2 in the positive set. Finally, experimentally rejected NMV_141 and NMV_227 were given an extra weight of 1 and 2 respectively in the negative set.

The final SVM classifier was built on a rbf kernel considering 11 features whose Spearman correlation coefficient is under 0.55 that are: the two Nardini features; positive-negative Z-score, positive-positive Z-score, and 9 CIDER features; proportion of chain expanding residues, fraction of E, V, N residues, E versus D ratio, fraction of aromatic, Y, and hydrophobic residues, and K versus R ratio.

For comparison we also tested MolPhase[27] and ParSe v2[26], two tools that predict phase separation scaffold proteins.

Figures were drawn on R v. 4.2.1 (https://www.R-project.org/) and the ggplot2 package[63] UMAPs were drows with the uwot package (https://doi.org/10.48550/arXiv.1802.03426)

Data and code availability

All the codes used for the bioinformatic analysis and for the VFCpredict tool developed in this study are available at https://src.koda.cnrs.fr/igs/vfcpredict.

References

1 Zhang, X., Zheng, R., Li, Z. & Ma, J. Liquid-liquid Phase Separation in Viral Function. *J Mol Biol* **435**, 167955 (2023). https://doi.org:10.1016/j.jmb.2023.167955

2 Charman, M. *et al.* A viral biomolecular condensate coordinates assembly of progeny particles. *Nature* **616**, 332-338 (2023). https://doi.org:10.1038/s41586-023-05887-y

662 3 Koonin, E. V. *et al.* Global Organization and Proposed Megataxonomy of the Virus World.
663 *Microbiol Mol Biol Rev* **84** (2020). https://doi.org:10.1128/MMBR.00061-19
664 4 Koonin, E. V., Kuhn, J. H., Dolja, V. V. & Krupovic, M. Megataxonomy and global ecology of the
665 virosphere. *ISME J* **18** (2024). https://doi.org:10.1093/ismejo/wrad042
666 5 Homola, M. *et al.* Structure and replication cycle of a virus infecting climate-modulating alga
667 Emiliania huxleyi. *Sci Adv* **10**, eadk1954 (2024). https://doi.org:10.1126/sciadv.adk1954
668 6 Kuznetsov, Y. G., Klose, T., Rossmann, M. & McPherson, A. Morphogenesis of Mimivirus and Its
669 Viral Factories: an Atomic Force Microscopy Study of Infected Cells. *Journal of Virology* **87**,
670 11200-11213 (2013). https://doi.org:10.1128/Jvi.01372-13
671 7 Philippe, N., Shukla, A., Abergel, C. & Bisio, H. Genetic manipulation of giant viruses and their
672 host, Acanthamoeba castellanii. *Nat Protoc* (2023). https://doi.org:10.1038/s41596-023-00910-y
673 8 Alempic, J. M. *et al.* Functional redundancy revealed by the deletion of the mimivirus GMC-
674 oxidoreductase genes. *Microlife* **5**, uqae006 (2024). https://doi.org:10.1093/femsml/uqae006
675 9 Bisio, H. *et al.* Evolution of giant pandoravirus revealed by CRISPR/Cas9. *Nat Commun* **14**, 428
676 (2023). https://doi.org:10.1038/s41467-023-36145-4
677 10 Liu, Y. *et al.* Virus-encoded histone doublets are essential and form nucleosome-like structures.
678 *Cell* **184**, 4237-4250 e4219 (2021). https://doi.org:10.1016/j.cell.2021.06.032
679 11 Fridmann-Sirkis, Y. *et al.* Efficiency in Complexity: Composition and Dynamic Nature of Mimivirus
680 Replication Factories. *J Virol* **90**, 10039-10047 (2016). https://doi.org:10.1128/JVI.01319-16
681 12 Bracha, D., Walls, M. T. & Brangwynne, C. P. Probing and engineering liquid-phase organelles.
682 *Nat Biotechnol* **37**, 1435-1445 (2019). https://doi.org:10.1038/s41587-019-0341-6
683 13 Hu, G. *et al.* flDPnn: Accurate intrinsic disorder prediction with putative propensities of disorder
684 functions. *Nat Commun* **12**, 4438 (2021). https://doi.org:10.1038/s41467-021-24773-7
685 14 Sharma, D., Coulibaly, F. & Kondabagil, K. Mimivirus encodes an essential MC1-like non-histone
686 architectural protein involved in DNA condensation. *bioRxiv*, 2024.2002.2022.580433 (2024).
687 https://doi.org:10.1101/2024.02.22.580433
688 15 Zarin, T. *et al.* Identifying molecular features that are associated with biological function of
689 intrinsically disordered protein regions. *Elife* **10** (2021). https://doi.org:ARTN e60220

690 10.7554/eLife.60220
691 16 Cohan, M. C., Shinn, M. K., Lalmansingh, J. M. & Pappu, R. V. Uncovering Non-random Binary
692 Patterns Within Sequences of Intrinsically Disordered Proteins. *Journal of Molecular Biology* **434**
693 (2022). https://doi.org:ARTN 167373

694 10.1016/j.jmb.2021.167373
695 17 Wang, J. *et al.* A Molecular Grammar Governing the Driving Forces for Phase Separation of Prion-
696 like RNA Binding Proteins. *Cell* **174**, 688-699 e616 (2018).
697 https://doi.org:10.1016/j.cell.2018.06.006
698 18 Lyons, H. *et al.* Functional partitioning of transcriptional regulators by patterned charge blocks.
699 *Cell* **186**, 327-+ (2023). https://doi.org:10.1016/j.cell.2022.12.013
700 19 Greig, J. A. *et al.* Arginine-Enriched Mixed-Charge Domains Provide Cohesion for Nuclear Speckle
701 Condensation. *Mol Cell* **77**, 1237-1250 e1234 (2020).
702 https://doi.org:10.1016/j.molcel.2020.01.025
703 20 Patil, A. *et al.* A disordered region controls cBAF activity via condensation and partner
704 recruitment. *Cell* **186**, 4936-4955 e4926 (2023). https://doi.org:10.1016/j.cell.2023.08.032
705 21 Boija, A. *et al.* Transcription Factors Activate Genes through the Phase-Separation Capacity of
706 Their Activation Domains. *Cell* **175**, 1842-+ (2018). https://doi.org:10.1016/j.cell.2018.10.042
707 22 King, M. R. *et al.* Macromolecular condensation organizes nucleolar sub-phases to set up a pH
708 gradient. *Cell* **187**, 1889-1906 e1824 (2024). https://doi.org:10.1016/j.cell.2024.02.029

709 23 Aylward, F. O., Moniruzzaman, M., Ha, A. D. & Koonin, E. V. A phylogenomic framework for
710 charting the diversity and evolution of giant viruses. *PLoS Biol* **19**, e3001430 (2021).
711 https://doi.org:10.1371/journal.pbio.3001430
712 24 Rigou, S., Santini, S., Abergel, C., Claverie, J.-M. & Legendre, M. (Research Square, 2022).
713 25 Gaïa, M. *et al.* Egoviruses: distant relatives of poxviruses abundant in the gut microbiome of
714 humans and animals worldwide. *bioRxiv*, 2024.2003.2023.586382 (2024).
715 https://doi.org:10.1101/2024.03.23.586382
716 26 Ibrahim, A. Y. *et al.* Intrinsically disordered regions that drive phase separation form a robustly
717 distinct protein class. *J Biol Chem* **299** (2023). https://doi.org:ARTN 102801

718 10.1016/j.jbc.2022.102801
719 27 Liang, Q. *et al.* MolPhase, an advanced prediction algorithm for protein phase separation. *EMBO*
720 *J* **43**, 1898-1918 (2024). https://doi.org:10.1038/s44318-024-00090-9
721 28 Schulz, F., Abergel, C. & Woyke, T. Giant virus biology and diversity in the era of genome-
722 resolved metagenomics. *Nat Rev Microbiol* (2022). https://doi.org:10.1038/s41579-022-00754-5
723 29 Das, R. K. & Pappu, R. V. Conformations of intrinsically disordered proteins are influenced by
724 linear sequence distributions of oppositely charged residues. *Proc Natl Acad Sci U S A* **110**,
725 13392-13397 (2013). https://doi.org:10.1073/pnas.1304749110
726 30 Borden, K. L. B. & Volpon, L. The diversity, plasticity, and adaptability of cap-dependent
727 translation initiation and the associated machinery. *RNA Biol* **17**, 1239-1251 (2020).
728 https://doi.org:10.1080/15476286.2020.1766179
729 31 Raoult, D. *et al.* The 1.2-megabase genome sequence of Mimivirus. *Science* **306**, 1344-1350
730 (2004). https://doi.org:10.1126/science.1101485
731 32 Joklik, W. K. & Becker, Y. The Replication and Coating of Vaccinia DNA. *J Mol Biol* **10**, 452-474
732 (1964). https://doi.org:10.1016/s0022-2836(64)80066-8
733 33 Kamagata, K. *et al.* Molecular principles of recruitment and dynamics of guest proteins in liquid
734 droplets. *Sci Rep* **11**, 19323 (2021). https://doi.org:10.1038/s41598-021-98955-0
735 34 Smug, B. J., Szczepaniak, K., Rocha, E. P. C., Dunin-Horkawicz, S. & Mostowy, R. J. Ongoing
736 shuffling of protein fragments diversifies core viral functions linked to interactions with bacterial
737 hosts. *Nat Commun* **14**, 7460 (2023). https://doi.org:10.1038/s41467-023-43236-9
738 35 Storz, J. F. Causes of molecular convergence and parallelism in protein evolution. *Nat Rev Genet*
739 **17**, 239-250 (2016). https://doi.org:10.1038/nrg.2016.11
740 36 Tolonen, N., Doglio, L., Schleich, S. & Krijnse Locker, J. Vaccinia virus DNA replication occurs in
741 endoplasmic reticulum-enclosed cytoplasmic mini-nuclei. *Mol Biol Cell* **12**, 2031-2046 (2001).
742 https://doi.org:10.1091/mbc.12.7.2031
743 37 Boyle, K. A., Greseth, M. D. & Traktman, P. Genetic Confirmation that the H5 Protein Is Required
744 for Vaccinia Virus DNA Replication. *Journal of Virology* **89**, 6312-6327 (2015).
745 https://doi.org:10.1128/Jvi.00445-15
746 38 Kay, N. E. *et al.* Biochemical and Biophysical Properties of a Putative Hub Protein Expressed by
747 Vaccinia Virus. *J Biol Chem* **288**, 11470-11481 (2013). https://doi.org:10.1074/jbc.M112.442012
748 39 Huang, Y. *et al.* Multi-omics characterization of the monkeypox virus infection. *Nat Commun* **15**,
749 6778 (2024). https://doi.org:10.1038/s41467-024-51074-6
750 40 Beaud, G. & Beaud, R. Preferential virosomal location of underphosphorylated H5R protein
751 synthesized in vaccinia virus-infected cells. *J Gen Virol* **78 (Pt 12)**, 3297-3302 (1997).
752 https://doi.org:10.1099/0022-1317-78-12-3297
753 41 Greseth, M. D. & Traktman, P. The Life Cycle of the Vaccinia Virus Genome. *Annu Rev Virol* **9**,
754 239-259 (2022). https://doi.org:10.1146/annurev-virology-091919-104752

755 42 Mallardo, M. *et al.* Relationship between vaccinia virus intracellular cores, early mRNAs, and
756 DNA replication sites. *Journal of Virology* **76**, 5167-5183 (2002).
757 https://doi.org:10.1128/Jvi.76.10.5167-5183.2002
758 43 Rodrigues, R. A. L. *et al.* Analysis of a Marseillevirus Transcriptome Reveals Temporal Gene
759 Expression Profile and Host Transcriptional Shift. *Front Microbiol* **11**, 651 (2020).
760 https://doi.org:10.3389/fmicb.2020.00651
761 44 Fabre, E. *et al.* Noumeavirus replication relies on a transient remote control of the host nucleus.
762 *Nat Commun* **8**, 15087 (2017). https://doi.org:10.1038/ncomms15087
763 45 Guglielmini, J., Woo, A. C., Krupovic, M., Forterre, P. & Gaia, M. Diversification of giant and large
764 eukaryotic dsDNA viruses predated the origin of modern eukaryotes. *Proc Natl Acad Sci U S A*
765 **116**, 19585-19592 (2019). https://doi.org:10.1073/pnas.1912006116
766 46 Mutsafi, Y., Zauberman, N., Sabanay, I. & Minsky, A. Vaccinia-like cytoplasmic replication of the
767 giant Mimivirus. *Proc Natl Acad Sci U S A* **107**, 5978-5982 (2010).
768 https://doi.org:10.1073/pnas.0912737107
769 47 Katsafanas, G. C. & Moss, B. Colocalization of transcription and translation within cytoplasmic
770 poxvirus factories coordinates viral expression and subjugates host functions. *Cell Host &*
771 *Microbe* **2**, 221-228 (2007). https://doi.org:10.1016/j.chom.2007.08.005
772 48 Legendre, M. *et al.* Thirty-thousand-year-old distant relative of giant icosahedral DNA viruses
773 with a pandoravirus morphology. *Proc Natl Acad Sci U S A* **111**, 4274-4279 (2014).
774 https://doi.org:10.1073/pnas.1320670111
775 49 Legendre, M. *et al.* In-depth study of Mollivirus sibericum, a new 30,000-y-old giant virus
776 infecting Acanthamoeba. *Proc Natl Acad Sci U S A* **112**, E5327-5335 (2015).
777 https://doi.org:10.1073/pnas.1510795112
778 50 Bertaux, L., Lartigue, A. & Jeudy, S. Giant Mimiviridae CsCl Purification Protocol. *Bio Protoc* **10**,
779 e3827 (2020). https://doi.org:10.21769/BioProtoc.3827
780 51 Bisio, H., Krishnan, A., Marq, J. B. & Soldati-Favre, D. Toxoplasma gondii phosphatidylserine
781 flippase complex ATP2B-CDC50.4 critically participates in microneme exocytosis. *PLoS Pathog* **18**,
782 e1010438 (2022). https://doi.org:10.1371/journal.ppat.1010438
783 52 Casabona, M. G., Vandenbrouck, Y., Attree, I. & Coute, Y. Proteomic characterization of
784 Pseudomonas aeruginosa PAO1 inner membrane. *Proteomics* **13**, 2419-2423 (2013).
785 https://doi.org:10.1002/pmic.201200565
786 53 Perez-Riverol, Y. *et al.* The PRIDE database resources in 2022: a hub for mass spectrometry-
787 based proteomics evidences. *Nucleic Acids Res* **50**, D543-D552 (2022).
788 https://doi.org:10.1093/nar/gkab1038
789 54 Bouyssie, D. *et al.* Proline: an efficient and user-friendly software suite for large-scale
790 proteomics. *Bioinformatics* **36**, 3148-3155 (2020).
791 https://doi.org:10.1093/bioinformatics/btaa118
792 55 Coute, Y., Bruley, C. & Burger, T. Beyond Target-Decoy Competition: Stable Validation of Peptide
793 and Protein Identifications in Mass Spectrometry-Based Discovery Proteomics. *Anal Chem* **92**,
794 14898-14906 (2020). https://doi.org:10.1021/acs.analchem.0c00328
795 56 Wieczorek, S. *et al.* DAPAR & ProStaR: software to perform statistical analyses in quantitative
796 discovery proteomics. *Bioinformatics* **33**, 135-136 (2017).
797 https://doi.org:10.1093/bioinformatics/btw580
798 57 Necci, M., Piovesan, D., Dosztanyi, Z. & Tosatto, S. C. E. MobiDB-lite: fast and highly specific
799 consensus prediction of intrinsic disorder in proteins. *Bioinformatics* **33**, 1402-1404 (2017).
800 https://doi.org:10.1093/bioinformatics/btx015

801 58 Holehouse, A. S., Das, R. K., Ahad, J. N., Richardson, M. O. & Pappu, R. V. CIDER: Resources to
802 Analyze Sequence-Ensemble Relationships of Intrinsically Disordered Proteins. *Biophys J* **112**, 16-
803 21 (2017). https://doi.org:10.1016/j.bpj.2016.11.3200
804 59 LeCun, Y., Bottou, L., Orr, G. B. & Müller, K. R. Efficient backprop. *Lect Notes Comput Sc* **1524**, 9-
805 50 (1998). https://doi.org:Doi 10.1007/3-540-49430-8_2
806 60 Steinegger, M. & Soding, J. MMseqs2 enables sensitive protein sequence searching for the
807 analysis of massive data sets. *Nat Biotechnol* **35**, 1026-1028 (2017).
808 https://doi.org:10.1038/nbt.3988
809 61 Notredame, C., Higgins, D. G. & Heringa, J. T-Coffee: A novel method for fast and accurate
810 multiple sequence alignment. *J Mol Biol* **302**, 205-217 (2000).
811 https://doi.org:10.1006/jmbi.2000.4042
812 62 Eddy, S. R. A new generation of homology search tools based on probabilistic inference. *Genome*
813 *Inform* **23**, 205-211 (2009).
814 63 Villanueva, R. A. M. & Chen, Z. J. ggplot2: Elegant Graphics for Data Analysis, 2nd edition. *Meas-*
815 *Interdiscip Res* **17**, 160-167 (2019). https://doi.org:10.1080/15366367.2019.1565254

816

817

Figures

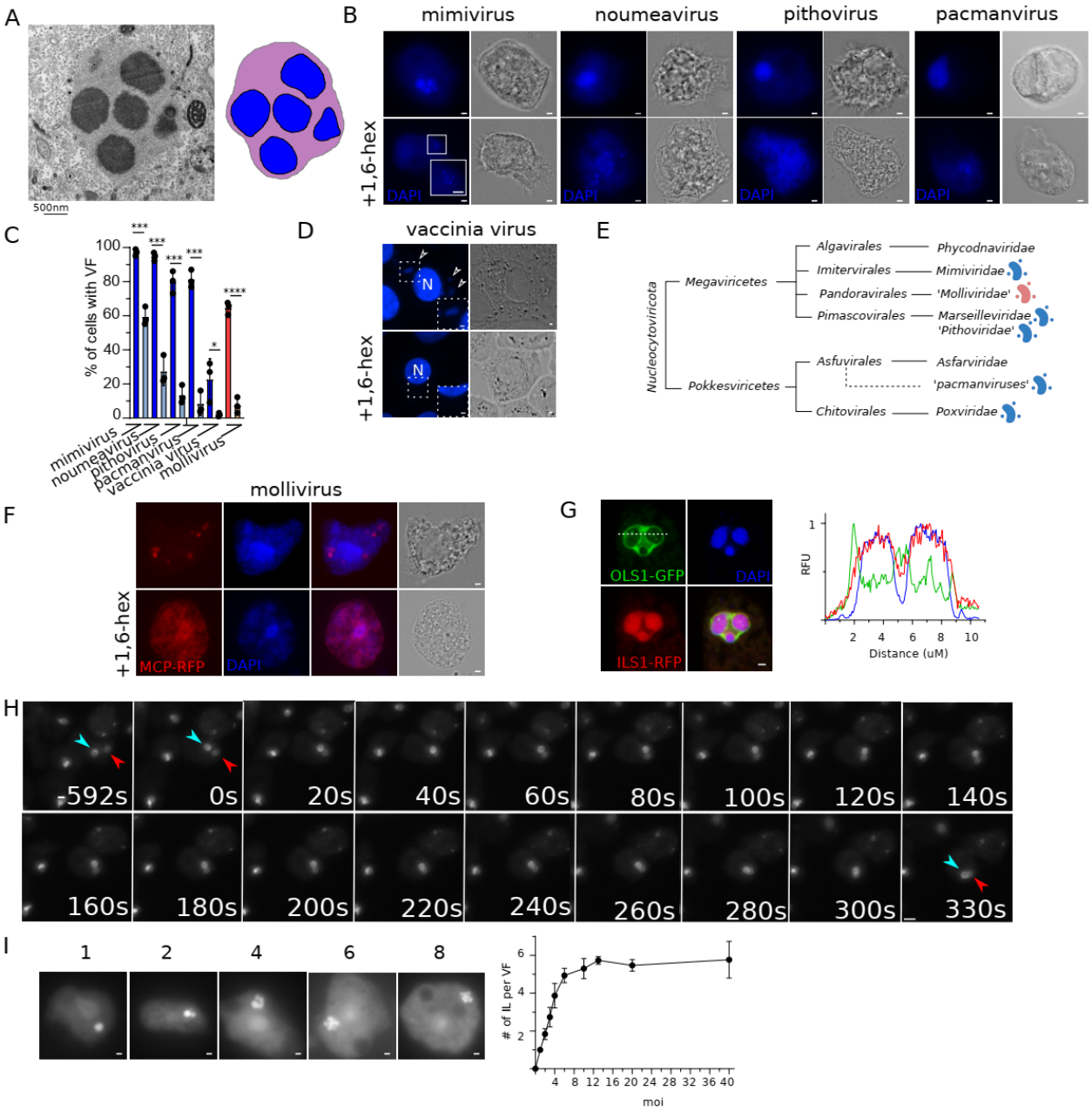


Figure 1. *Nucleocytoviricota* viral factories (VFs) are biomolecular condensates.

(A) Negative staining electron microscopy imaging of an ultrathin section of infected *A. castellanii* cell with mimivirus VF formed in the cytoplasm. Image was acquired 6h post infection (pi) at a MOI=20. Scale bar: 500nm. A cartoon representing the two layers of the viral factory is also shown. Inner layer (IL) is shown in blue while outer layer (OL) is shown in purple.

(B) Representative light fluorescence microscopy images of *A. castellanii* cells infected with different viruses belonging to *Nucleocytoviricota*. VFs were labelled using DAPI and treatment with 10% 1,6-hexanediol was performed for ten minutes after 3h, 2h, 6h and 6h pi for mimivirus, noumeavirus, pithovirus and pacmanvirus, respectively. Scale bar: 1µm.

(C) Quantification of the experiments shown in Figure 1B, Figure 1D and Figure 1F. Data correspond to the mean ± SD of 3 independent experiments. Quantification performed base on DAPI staining is shown in blue, while quantification performed base on mollivirus MCP-RFP is shown in red. ns ($P > 0.05$), * ($P \leq 0.05$), ** ($P \leq 0.01$), *** ($P \leq 0.001$) and **** ($P \leq 0.0001$).

(D) Representative light fluorescence microscopy images of Vero cells infected with vaccinia virus. VFs were labelled using DAPI and treatment with 10% 1,6-hexanediol was performed for ten minutes after 2hpi. Scale bar: 1µm. Unfilled arrowhead indicate the VFs while the nucleus of the host cell is highlighted with and N. A zoom of the perinuclear zone is shown in the inset.

(E) Taxonomy of viruses belonging to the phylum *Nucleocytoviricota*. Families with a member where 1,6-hexanediol dissolved their VF (blue) or the proposed "Virion Factory" (red) are indicated. Image was adapted from [4].

(F) Representative light fluorescence microscopy images of *A. castellanii* cells expressing C-terminally RFP tagged MCP. Representative images of cells infected with mollivirus are shown and treatment with 10% 1,6-hexanediol was performed for ten minutes 6h pi. Scale bar: 1µm.

(G) *A. castellanii* cells expressing C-terminally tagged OLS1-GFP (R561) and ILS1-RFP (R252) were infected with mimivirus. OLS1-GFP localized to the OL of the VF and ILS1-RFP to the IL of the VF. DAPI: DNA. Scale bar: 1µm. Line profiles (right) corresponding to the white dashed line show fluorescence patterns.

(H) Live-cell imaging of mimivirus infected-*A. castellanii* expressing OLS1-GFP as a marker of the OL of the VF. Infection was allowed to proceed for 3h and recording was performed every 2s. Two viral factories (marked with magenta and red arrowheads) fused their OL during the recording and drifted together for the entire time of recording (1 hour and 30 minutes). Scale bar: 10µm.

(I) Representative light fluorescence microscopy images of *A. castellanii* cells infected with mimivirus harboring VF with different numbers of ILs. Quantification of the number of ILs present in a VF in function of the MOI is shown on the right. Data correspond to the mean ± SD of 3 independent experiments. Scale bar: 1µm.

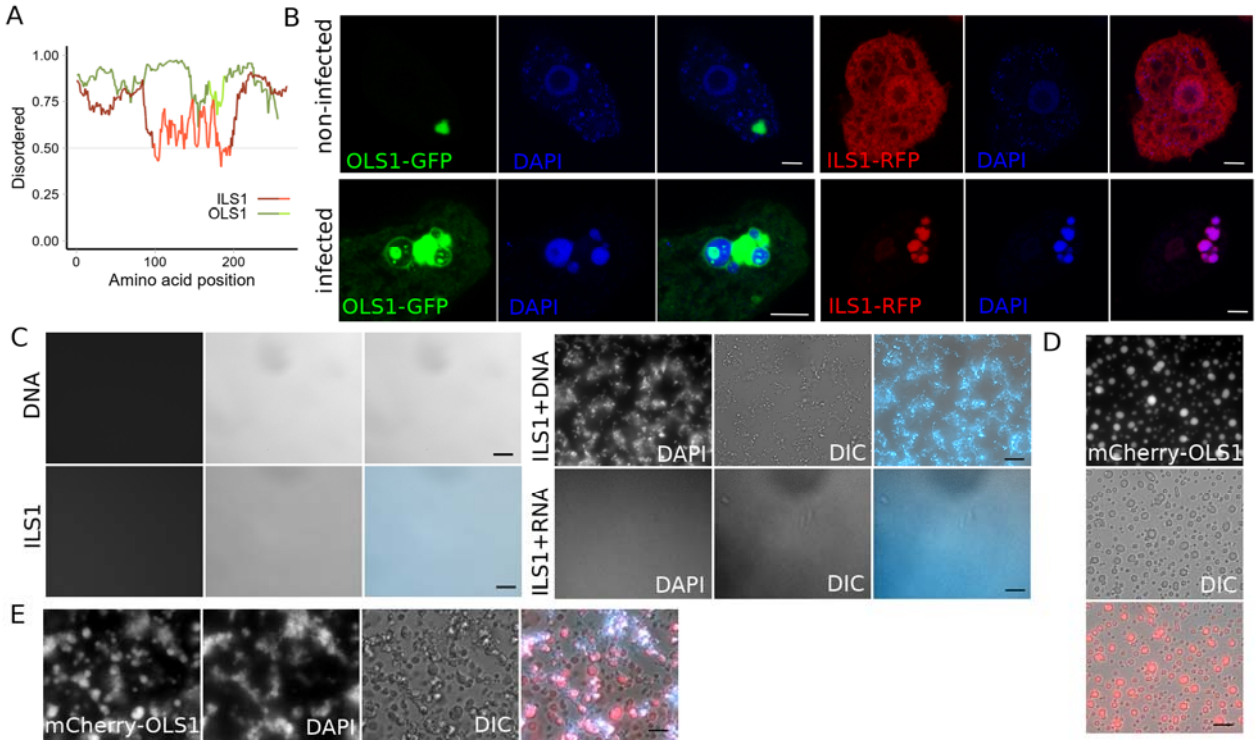


Figure 2. ILS1 and OLS1 are the scaffold proteins of each layer of the VF

(A) Predicted disorder tendency of OLS1 (green) and ILS1 (red).). Considered disordered regions are shown in darker color as predicted by MobiDB-lite while the numeric value was calculated by IUPred to confirm the prediction.

(B) *A. castellanii* cells expressing C-terminally tagged OLS1-GFP or ILS1-RFP were infected or not with mimivirus. In absence of infection OLS1-GFP shows signs of PS at the cytoplasm of the amoeba while ILS1-RFP is shown diffused in the cytoplasm and nucleus. Upon infection, OLS1-GFP localized to the OL of the VF and ILS1-RFP to the IL of the VF. DAPI: DNA. Scale bar: 5µm.

(C) *In vitro* PS of ILS1 in presence or absence of DNA or RNA. ILS1 was used at 5 µM, DNA at 10 µg/mL and RNA at 100 µg/mL. DAPI was used to confirm co-PS of protein and nucleic acid. Scale bar: 10µm.

(D) *In vitro* PS of mCherry-OLS1 at 50mM NaCl. OLS1 was used at 5µM. Scale bar: 10µm.

(E) *In vitro* PS of mCherry-OLS1 and ILS1 in presence of 10 µg/mL DNA. Both proteins were used at 5 µM. DAPI was used to confirm co-PS of protein and nucleic acid. Scale bar: 10µm.

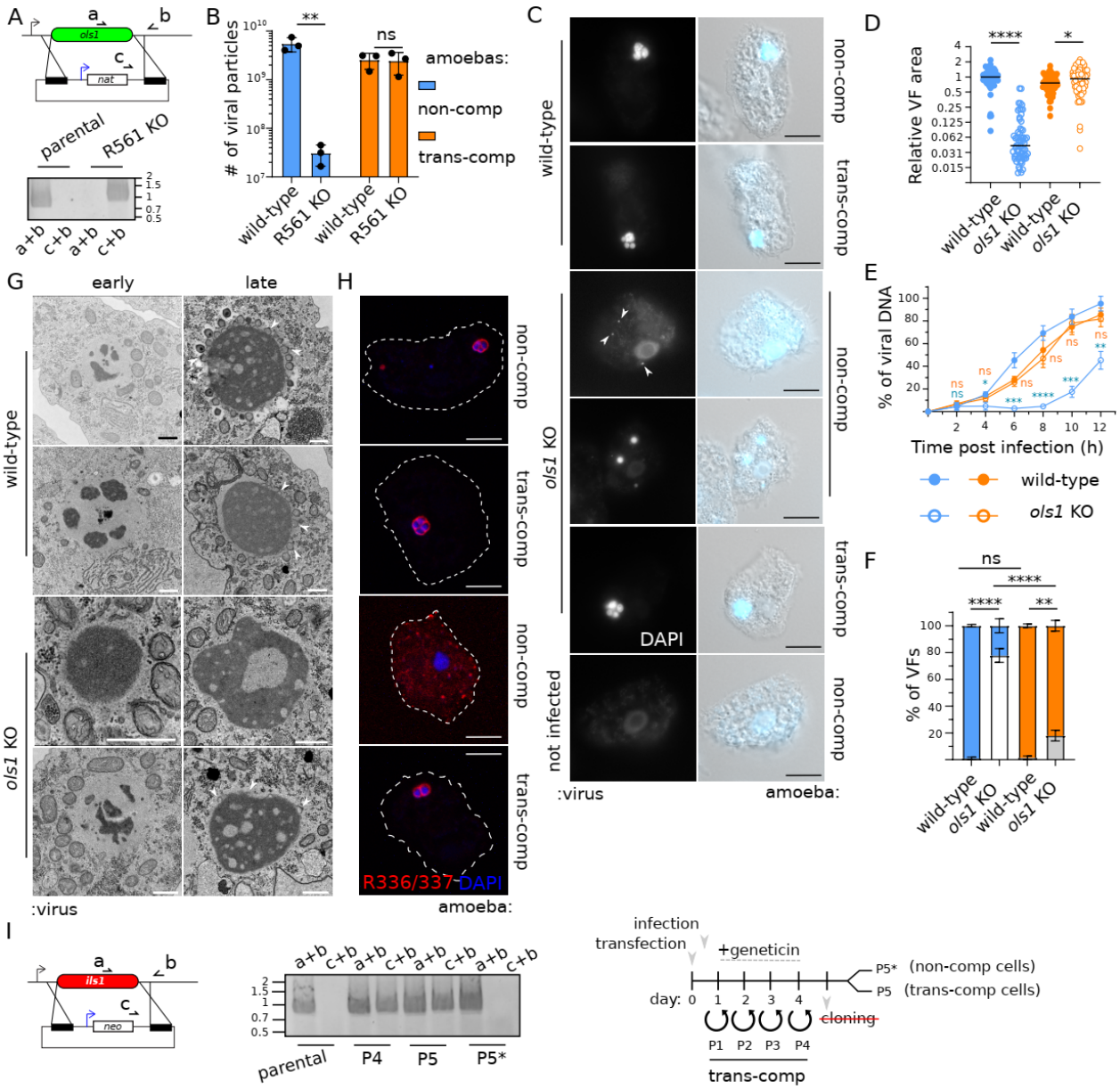


Figure 3. ILS1 and OLS1 play key roles in the mimivirus infection cycle

(A) Schematic representation of the vector and strategy used for *ols1* KO. Selection cassette was introduced by homologous recombination and recombinant viruses were generated, selected and cloned. *nat*: Nourseothricin N-acetyl transferase. Primers annealing locations

are shown and successful KO and clonality is demonstrated by PCR. Expected size: a+b: 850bp in parental locus. c+b: 990bp in recombinant locus.

(B) Quantification of the number of viruses produced upon knock-out of *ols1* in mimivirus. Infection was performed in non-complemented cells and in a trans-complementing line expressing OLS1. Data correspond to the mean ± SD of 3 independent experiments.

(C) Representative light fluorescence microscopy images of *A. castellanii* cells infected with wild-type mimivirus or *ols1* KO viruses 6h pi. Infection was performed in non-complemented cells and in a trans-complementing line expressing OLS1. VFs were labelled using DAPI Scale bar: 10µm.

(D) Quantification of the size of VF generated as shown in C. At least 50 VF were recorded per condition during 3 independent experiments and the area of DAPI staining was measured using ImageJ. Infection results are shown in blue or orange when generated on non-complemented cells or trans-complementing amoebas respectively.

(E) DNA replication was analyzed by qPCR. Viral DNA is represented as a percentage of total DNA in the sample. Data correspond to the mean ± SD of 3 independent experiments. Infection results are shown in blue or orange when generated on non-complemented cells or trans-complementing amoebas respectively.

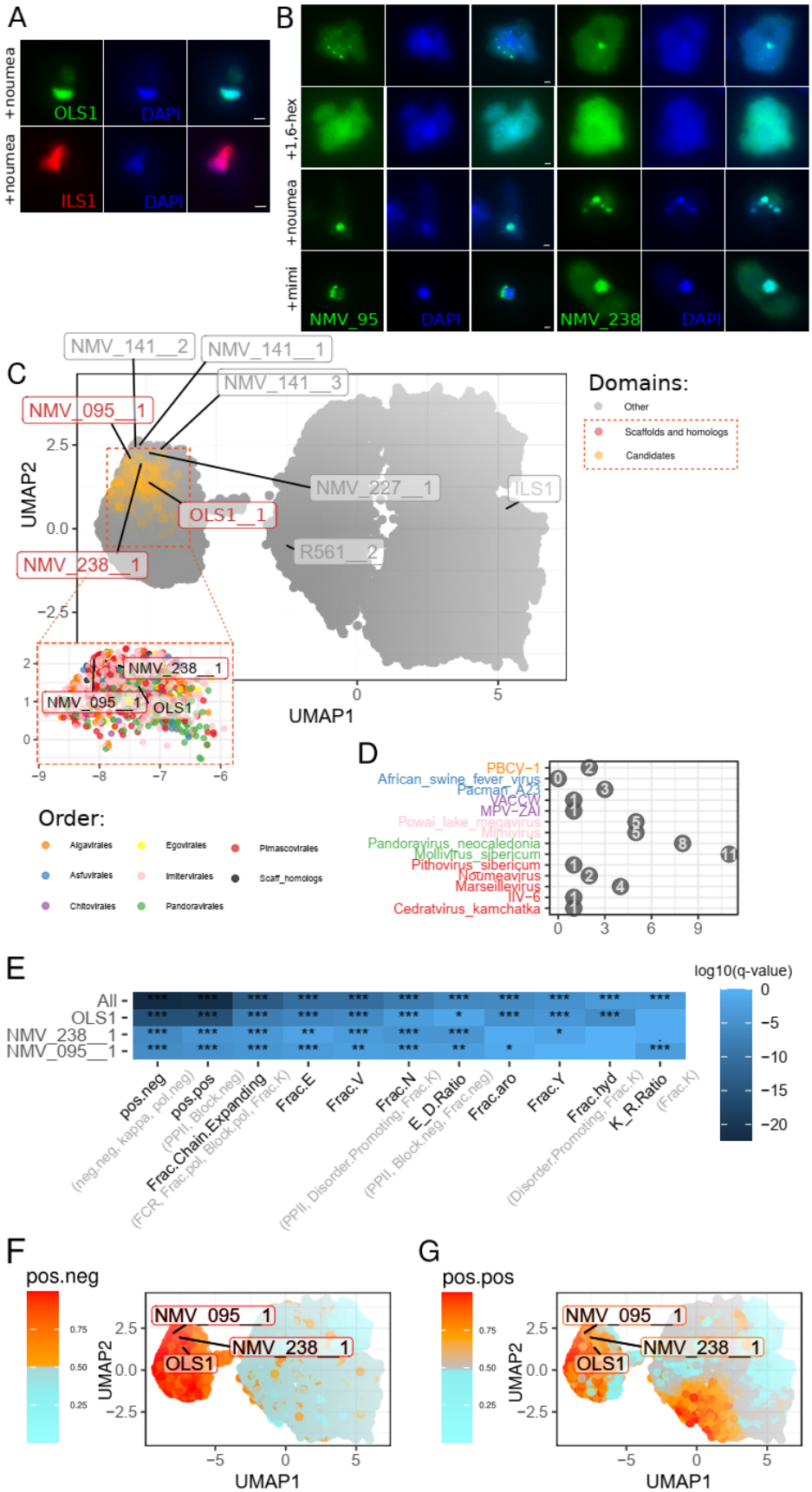
(F) Quantification of the number of VF either fused or separated in infected cells as shown in C. At least 100 infected cells were recorded per condition. Data correspond to the mean ± SD of 3 independent experiments. Fused VFs are shown in blue or orange when generated on non-complemented or trans-complementing amoebas respectively. Separated VFs are shown in white or grey when generated on wild-type cells or trans-complementing amoebas respectively.

(G) Electron microscopy imaging of the mimivirus replication cycle in *A. castellanii*. Images were acquired 4-6h pi and mimivirus particles (MOI=20) were used to infect non-complemented cells or cells expressing a copy of *ols1* (trans-complementing line). Nascent virions are indicated with white arrowheads. Scale bar: 1µm.

(H) Immunofluorescence demonstrating localization of client proteins from the OL of the VF. Proteins were endogenously tagged with 3xHA at the C-terminal and infection was carried out for 6h before fixation. VFs were labelled using DAPI. Scale bar: 1µm.

(I) Schematic representation of the strategy used to generate the recombinant. Selection cassette was introduced by homologous recombination and recombinant viruses were generated, selected and cloning attempted as indicated by the timeline. After passage 4, viruses were split and used to infect non-complemented amoebas (P5*) or trans-complementing amoebas (P5). Populations of recombinant viruses were followed by assessing integration of the selection cassette. Expected size: a+b: 870bp in parental locus. c+b: 890bp in recombinant locus.

ns ($P > 0.05$), * ($P \leq 0.05$), ** ($P \leq 0.01$), *** ($P \leq 0.001$) and **** ($P \leq 0.0001$).


A
+noumea
OLS1
DAPI
+noumea
ILS1
DAPI
B
+1,6-hex
+noumea
+mimi
NMV_95
DAPI
NMV_238
DAPI
C
NMV_141__2
NMV_141__1
NMV_141__3
NMV_095__1
NMV_227__1
ILS1
OLS1__1
R561__2
NMV_238__1
UMAP2
UMAP1
Domains:
Other
Scaffolds and homologs
Candidates
Order:
Algavirales
Egovirales
Pimascovirales
Asfuvirales
Imitervirales
Scaff_homologs
Chitovirales
Pandoravirales
D
PBCV-1
African_swine_fever_virus
Pacman_A23
VACCW
MPV-ZAI
Powai_lake_megavirus
Mimivirus
Pandoravirus_neocaledonia
Mollivirus_sibericum
Pithovirus_sibericum
Noumeavirus
Marseilleviruses
IIV-6
Cedratvirus_kamchatka
E
All
OLS1
NMV_238__1
NMV_095__1
pos.neg
(neg.neg, kappa, pol.neg)
pos.pos
(PPII, Block.neg)
Frac.Chain.Expanding
(FCR, Frac.pol, Block.pol, Frac.K)
Frac.E
Frac.V
Frac.N
(PPII, Disorder.Promoting, Frac.K)
E_D.Ratio
(PPII, Block.neg, Frac.neg)
Frac.aro
Frac.Y
Frac.hyd
(Disorder.Promoting, Frac.K)
K_R.Ratio
(Frac.K)
log10(q-value)
F
pos.neg
NMV_095__1
NMV_238__1
OLS1
UMAP2
UMAP1
G
pos.pos
NMV_095__1
NMV_238__1
OLS1
UMAP2
UMAP1

920

Figure 4. Identification of scaffold proteins by defining their molecular grammar

(A) *A. castellanii* cells expressing C-terminally tagged mimivirus OLS1-GFP or mimivirus ILS1-RFP were infected with noumeavirus. Both proteins localized to the VF of noumeavirus. DAPI: DNA. Scale bar: 1µm.

(B) *A. castellanii* cells expressing C-terminally tagged NMV_095 or NMV238 were either infected with noumeavirus, mimivirus or treated with 10% 1,6-hexanediol. VFs were labelled using DAPI 2-3 h pi. Scale bar: 1µm.

(C) UMAP representation of the IDRs in representative genomes and metagenomics giant viruses based on the 11 features selected for the classifier (H) 29,677 points are drawn, including 53 IDRs corresponding to scaffolds and homologs (red) and 1083 candidates retrieved by the classifier (orange). 800 genomes of *Imitervirales* were removed because they were over-represented. Before filtering, there was 76,555 points including 2967 candidate IDRs.

(D) Candidate proteins for phase separation predicted by the classifier in the representative genomes.

(E) Final classifier features and their corrected p-value in both mimivirus and noumeavirus genomes (All) or only one of them (OLS1, NMV). All homologs were considered. Marker within the heatmap gives the level of significance: *** < 0.001, ** < 0.01, * < 0.05, . < 0.1. Features that are in parenthesis are above 0.55 correlated in the scaffold proteins and homologs.

(F-G) Details of the most discriminant features on the UMAP (C). Pos.neg and pos.pos refers to the segregation of positive from negative residues or of positive residues to the other residues. Frac. goes for “fraction” and the chain expanding residues are E, D, R, K and P. FCR goes for “Fraction of Charged Residues”. PPII is the propensity to form polyproline II conformations. Disorder promoting gives the fraction of disordered promoting residues (including E). “pol” goes for polar residues, “hyd” for hydrophobic and finally, kappa is a measure of segregation of charges.

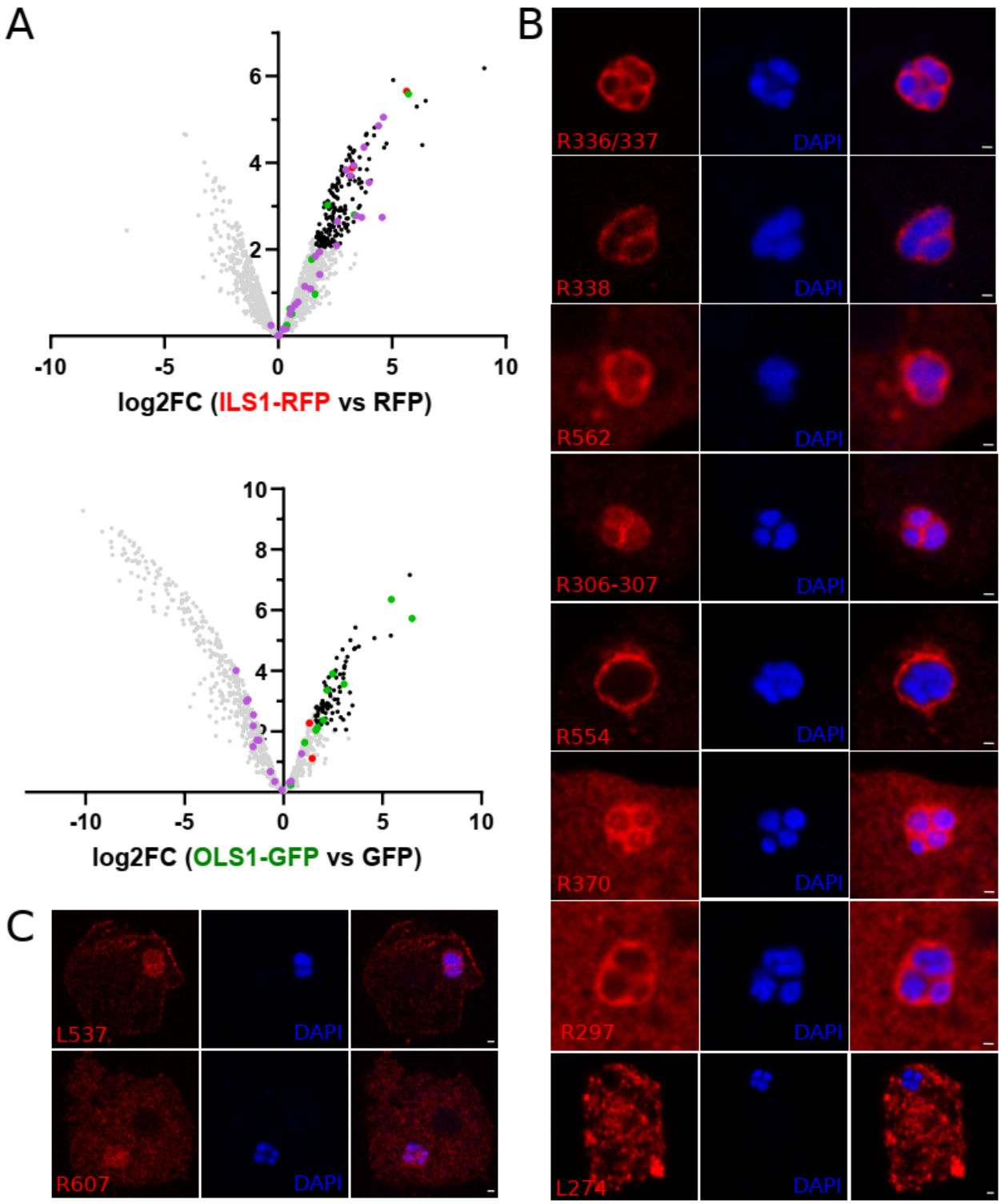


Figure 5. Identification of client proteins from the VF

(A) Co-immunoprecipitation (IP) experiments in *A. castellanii* cells expressing ILS1-RFP or OLS1-GFP and infected by mimivirus. Cells expressing RFP or GFP were utilized as

controls. Immunoprecipitated proteins were analyzed through MS-based label-free quantitative proteomics (three replicates per condition). The volcano plots represent the -log10 (limma p-value) on y axis plotted against the log2(Fold Change bait vs control) on x axis for each quantified protein (upper panel: OLS1-GFP versus GFP, bottom panel: ILS1-RFP versus RFP). Each dot represents a protein. Proteins with log2 (FoldChange) ≥ 1.6 (FoldChange > 3) and -log10(p-value) ≥ 2 (p-value = 0.01) compare to controls, were considered significant (Benjamini-Hochberg FDR < 2%) and are shown in black. Proteins confirmed to localize at the IL, OL or virions are shown in red, green or violet respectively. Detailed data are presented in Supplementary Table 3.

(B) Immunofluorescence demonstrating localization of proteins enriched in A. Proteins were endogenously tagged with 3xHA at the C-terminal and infection was carried out for 6 hours before fixation. VFs were labelled using DAPI. Scale bar: 1µm.

(C) Immunofluorescence demonstrating localization of proteins not enriched in A but still displaying a VF localization. Proteins were endogenously tagged with 3xHA at the C-terminal and infection was carried out for 6 hours before fixation. VFs were labelled using DAPI. Scale bar: 1µm.

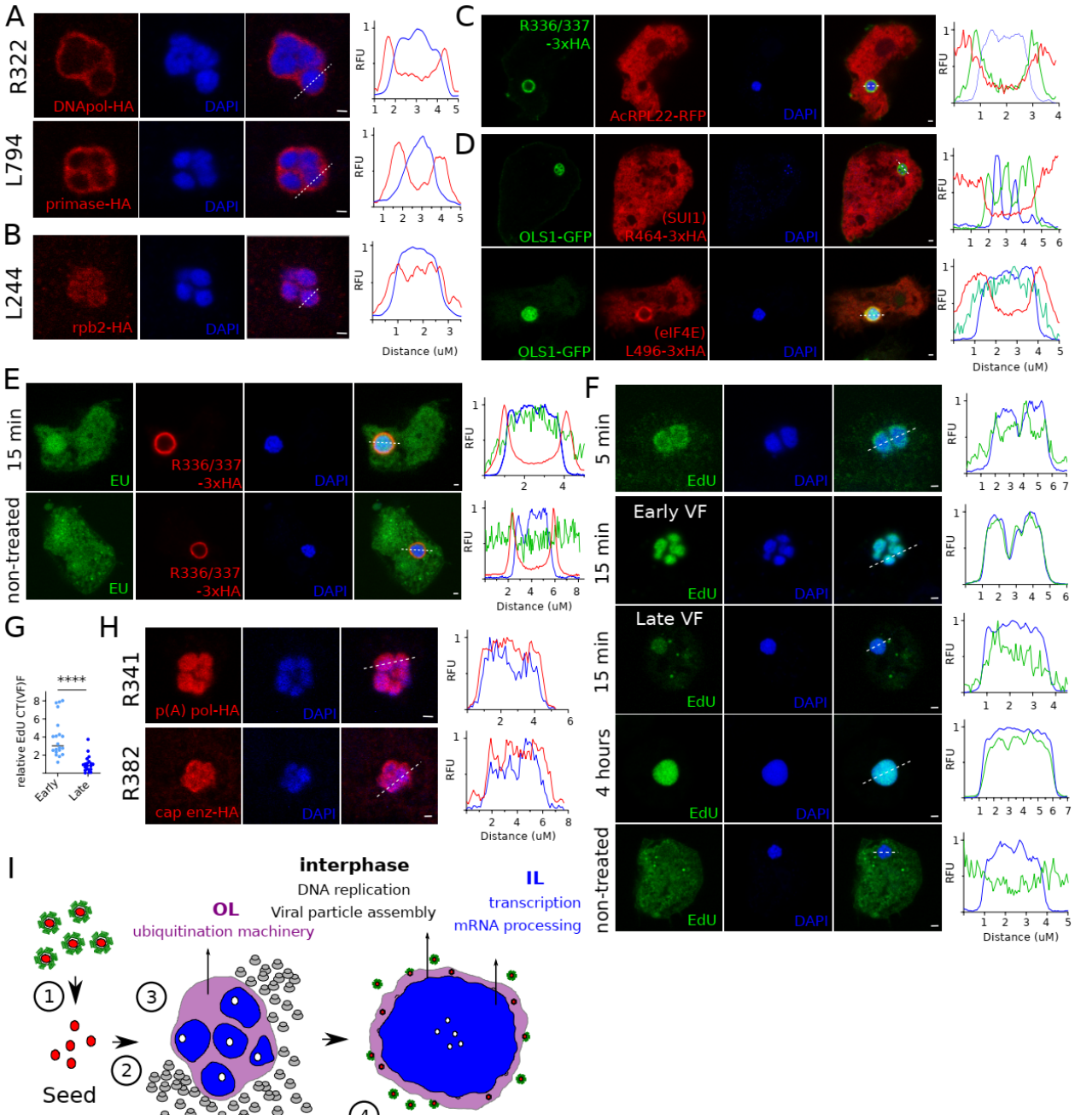


Figure 6. Sub-cellular and sub-organelle compartmentalization of the central dogma.

(A) Immunofluorescence demonstrating localization of proteins associated to DNA replication. Proteins were endogenously tagged with 3xHA at the C-terminal and infection was carried out for 6 hours before fixation. VFs were labelled using DAPI. Line profiles (bottom) corresponding to the white dashed line show fluorescence patterns. Scale bar: 1µm.

(B) Immunofluorescence demonstrating localization of the RNA polymerase subunit 2 (rpb2), associated to transcription. Protein was endogenously tagged with 3xHA at the C-terminal and infection was carried out for 6 hours before fixation. VFs were labelled using DAPI. Line profiles (bottom) corresponding to the white dashed line show fluorescence patterns. Scale bar: 1μm.

(C) Immunofluorescence demonstrating localization of host ribosomal protein RPL22-RFP, associated to translation. Protein is expressed from a second copy plasmid encoding *rpl22-rfp* and infection was carried out for 6 hours before fixation. IL and OL of the VFs were labelled using DAPI and R366/337-3xHA respectively. Line profiles (bottom) corresponding to the white dashed line show fluorescence patterns. Scale bar: 1μm.

(D) Immunofluorescence demonstrating localization of viral proteins associated to translation. Proteins were endogenously tagged with 3xHA at the C-terminal and infection was carried out for 6 hours before fixation. IL and OL of the VFs were labelled using DAPI and OLS1-GFP respectively. Line profiles (bottom) corresponding to the white dashed line show fluorescence patterns. Scale bar: 1μm.

(E) Detection of RNA synthesis by EU labelling. Viral infection by wild type viruses was allowed to proceed for 4-6 hours and labelling time is indicated. IL and OL of the VFs were labelled using DAPI and R366/337-3xHA respectively. Scale bar: 1μm.

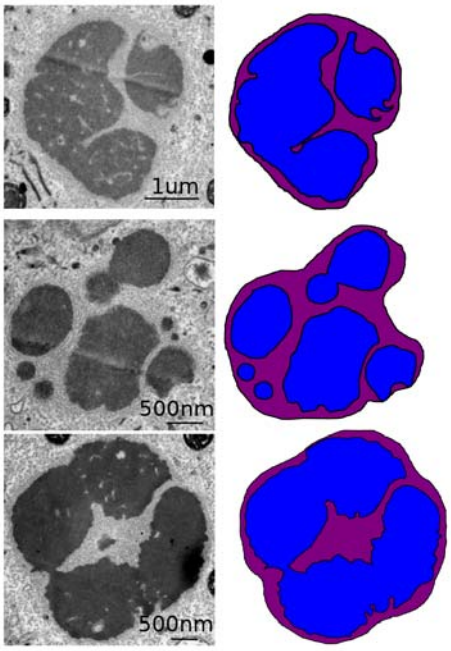
(F) Detection of DNA synthesis by EdU labelling. Viral infection by *thymidylate synthase* KO viruses was allowed to proceed for 4-6 hours and labelling time is indicated. IL of the VFs was labelled using DAPI. Scale bar: 1μm.

(G) Quantification of the corrected total VF fluorescence (CT(VF)F) of EdU labeling as shown in F. 20 VFs were recorder during 3 independent experiments and the intensity of EdU staining was measured using ImageJ. ns ($P > 0.05$), * ($P \leq 0.05$), ** ($P \leq 0.01$), *** ($P \leq 0.001$) and **** ($P \leq 0.0001$).

(H) Immunofluorescence demonstrating localization of proteins associated to mRNA maturation. Proteins were endogenously tagged with 3xHA at the C-terminal and infection was carried out for 6 hours before fixation. VFs were labelled using DAPI. Line profiles (bottom) corresponding to the white dashed line show fluorescence patterns. Scale bar: 1μm.

(I) Schematic depiction of the proposed model of mimivirus VF. Infective mimivirus (1) are internalized to the cell by phagocytosis and deliver an internal structure (core, red) containing the viral DNA (2). Individual viral factories are nucleated around the internalized cores (white). VFs are then generated by multilayered phase separation. Outer layer (OL, purple) quickly fuse in case of superinfection of host cells while Inner layer (IL, blue) fuse later. OL concentrates proteins associated to post-transcriptional regulation like ubiquitination while IL concentrates DNA as well as transcription and mRNA maturation. DNA synthesis and virion morphogenesis occur at the interface between OL and IL while glycosylation of virions (green) occurs upon exit of the VF. Early VFs (3) and late VFs (4) are depicted.

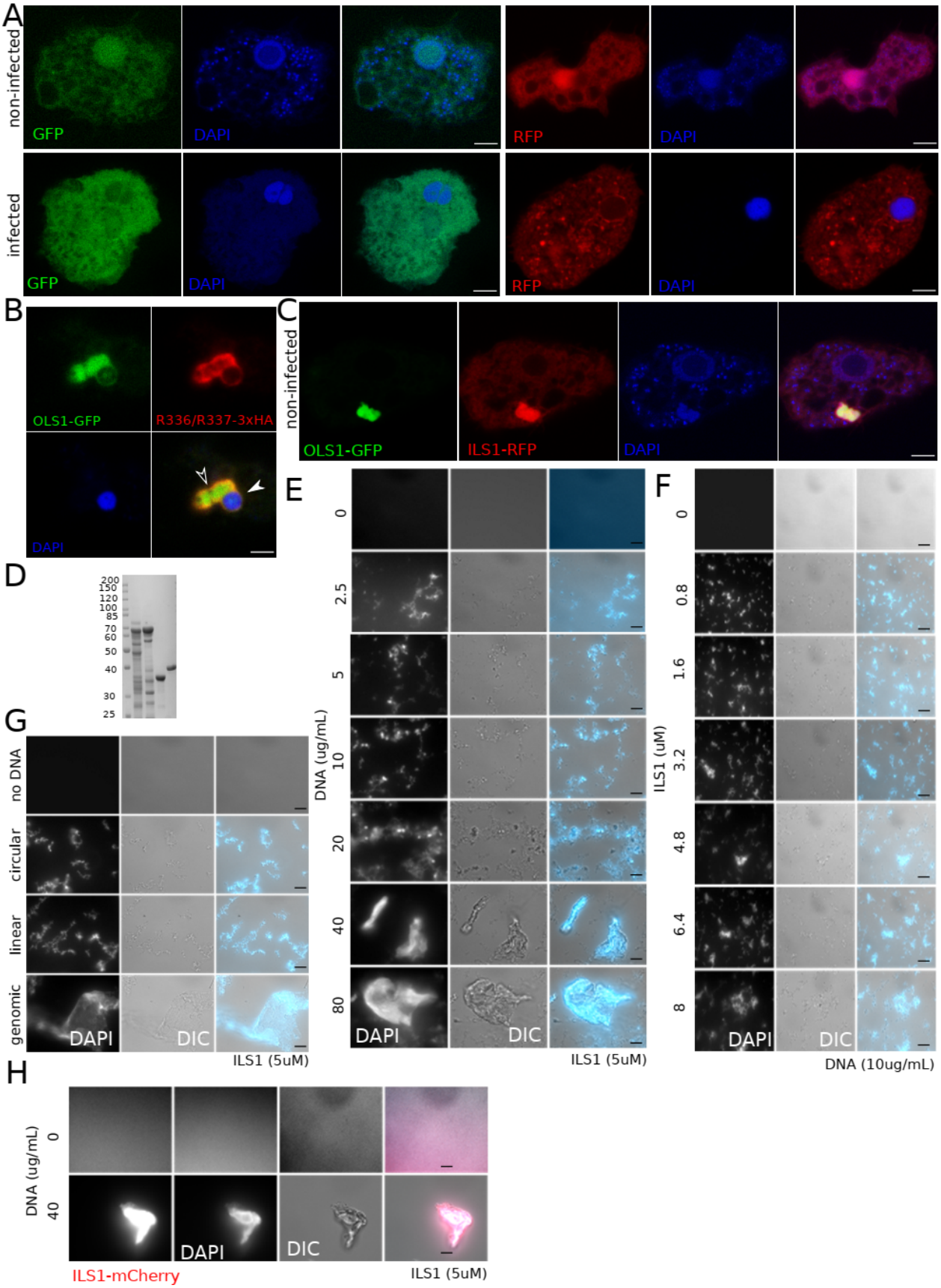
Supplemental Figures



Supplementary Figure 1.

(A) Electron microscopy imaging of the mimivirus VFs formed in the cytoplasm of *A. castellanii*. Potential coalescent events of the inner layer are highlighted. Image was acquired 6h pi at a MOI=20. Scale bar: 500nm. A cartoon representing the two layers of

the viral factory is also shown. Inner layer (IL) is shown in blue while outer layer (OL) is shown in purple.


A
non-infected
infected
GFP
DAPI
RFP
B
OLS1-GFP
R336/R337-3xHA
C
ILS1-RFP
D
200
150
120
100
85
70
60
50
40
30
25
E
DNA (ug/mL)
0
2.5
5
10
20
40
80
ILS1 (5uM)
F
ILS1 (uM)
0.8
1.6
3.2
4.8
6.4
8
DNA (10ug/mL)
G
no DNA
circular
linear
genomic
DIC
H
ILS1-mCherry

1024

Supplementary Figure 2.

(A) *A. castellanii* cells expressing GFP or RFP were infected or not with mimivirus. In absence of infection GFP and RFP are diffused in the cytoplasm and nucleus. Upon infection, no major changes in localization are detected. DAPI: DNA. Scale bar: 5 µm.

(B) Immunofluorescence demonstrating localization of client protein R336/R337-3xHA to the OL of the VF and OLS1-GFP BMC. Scale bar: 5 µm.

(C) *A. castellanii* cells expressing C-terminally tagged OLS1-GFP and ILS1-RFP illustrating that ILS1 is a client protein of OLS1. DAPI: DNA. Scale bar: 5 µm.

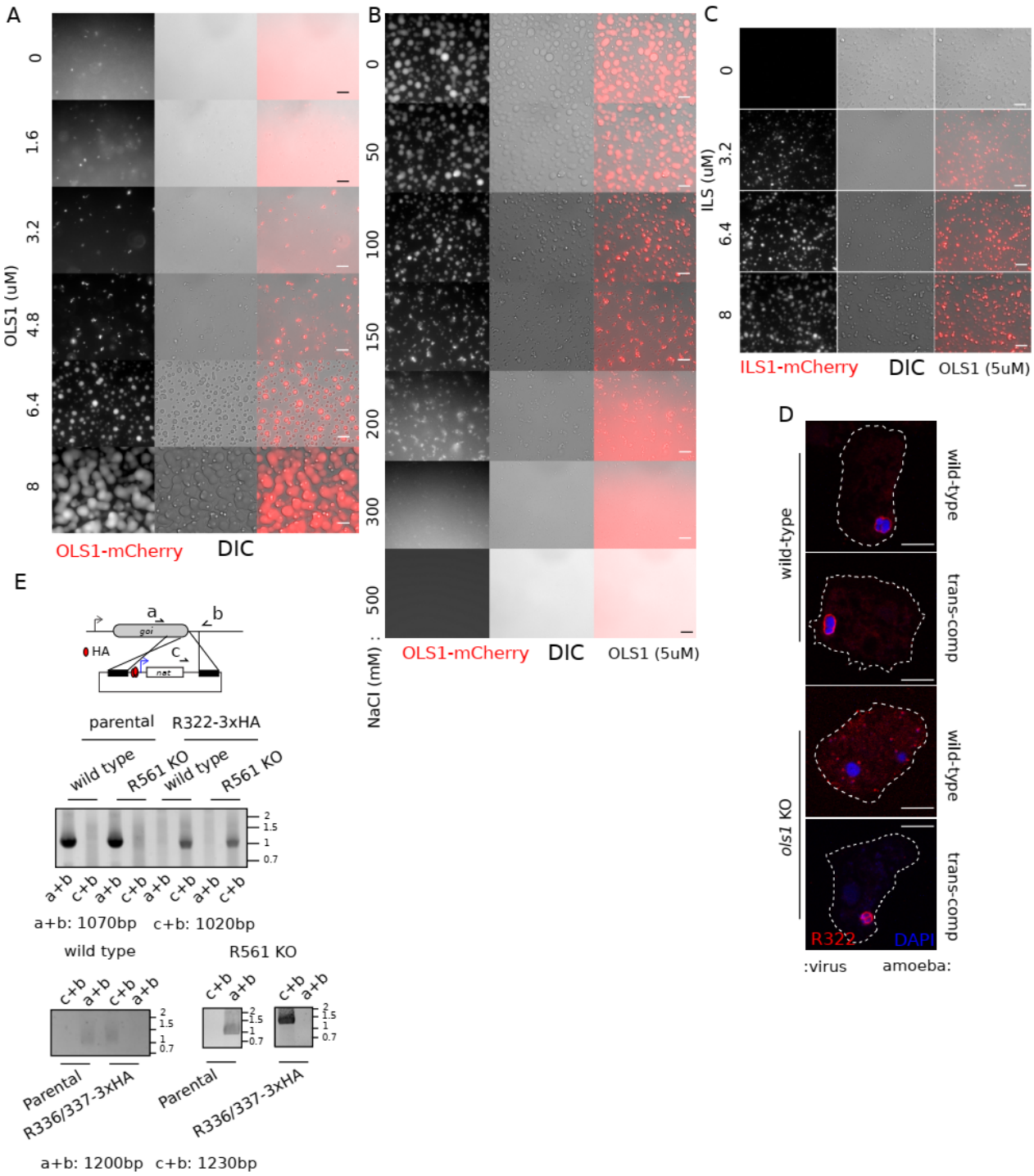
(D) Purified ILS1, OLS1, mCherry-ILS1 and mCherry-OLS1 proteins analysis on Sodium Dodecyl Sulphate-Polyacrylamide gel (SDS-PAGE). The mCherry fusion displayed significant contaminations with *E. coli* proteins but all results were confirmed with the non-tagged proteins. Loading order: mcherry-ILS1 (MW 69.5), mcherry-OLS1 (MW 70.2), V5-ILS1 (MW 30.8) and V5-OLS1 (MW 31.5). Proteins migrates at a highest molecular weight than predicted.

(E) *In vitro* PS of ILS1 in presence of DNA. ILS1 was used at 5 µM and DNA ranged from 0 to 80 µg/mL. DAPI was used to confirm co-PS between protein and nucleic acid. Scale bar: 10 µm.

(F) *In vitro* PS of ILS1 in presence of DNA. DNA was used at 10 µg/mL and ILS1 ranging from 0 to 8 µM. DAPI was used to confirm co-PS between protein and nucleic acid. Scale bar: 10 µm.

(G) *In vitro* PS of ILS1 in presence of DNA. ILS1 was used at 5 µM and DNA at 10 µg/mL. Circular and linear plasmid as well as mimivirus genomic DNA were compared. DAPI was used to confirm co-PS between protein and nucleic acid. Scale bar: 10 µm.

(H) *In vitro* PS of mCherry-ILS1 in presence of DNA. ILS1 was used at 5 µM and DNA at 40 µg/mL. Circular and linear plasmid as well as mimivirus genomic DNA were compared. DAPI was used to confirm co-PS between protein and nucleic acid. Scale bar: 10 µm.



Supplementary Figure 3.

(A) *In vitro* PS of mCherry-OLS1 at 50mM NaCl. mCherry-OLS1 was used at different concentrations. Scale bar: 10μm.

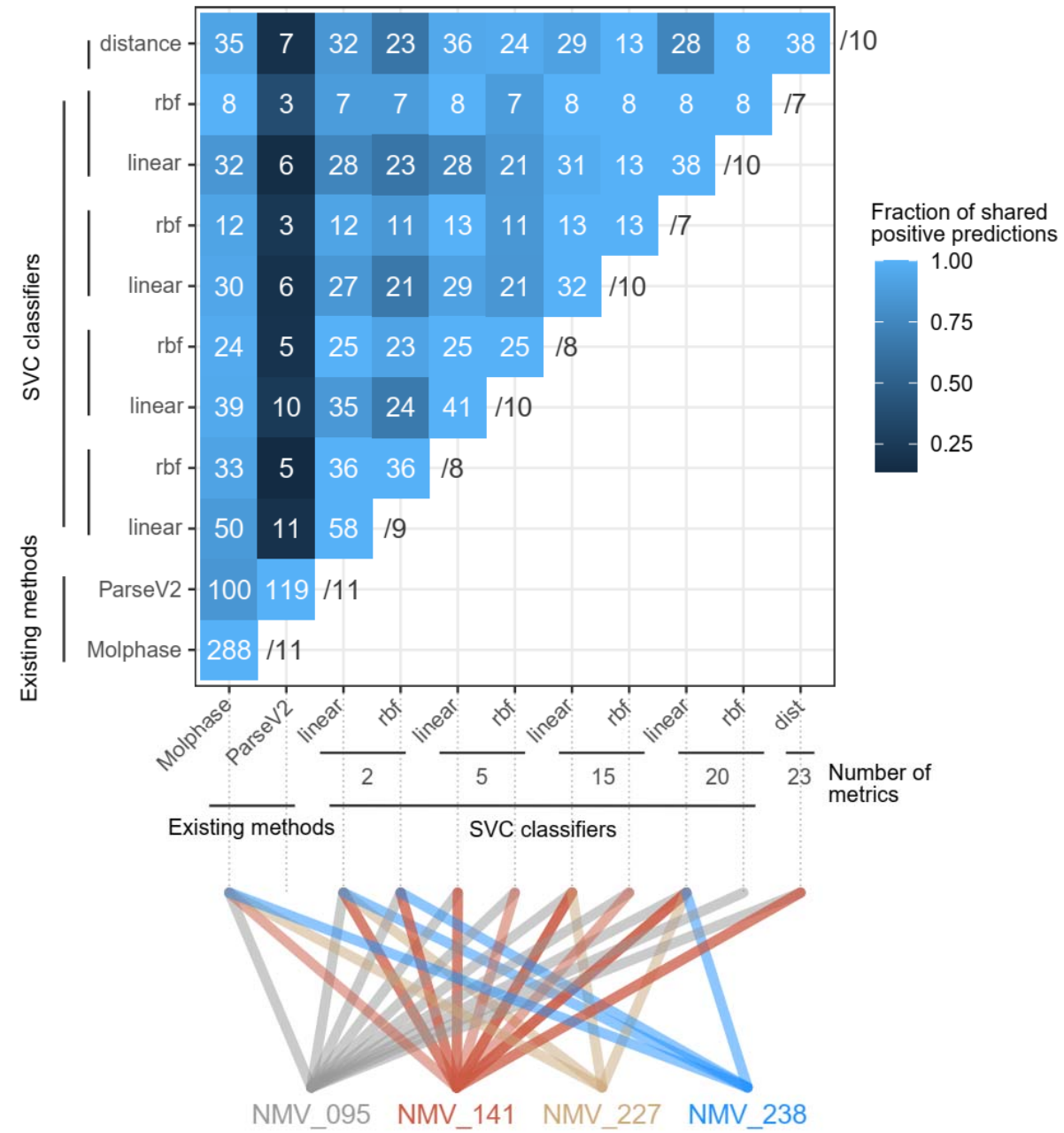
(B) *In vitro* PS of mCherry-OLS1 at different NaCl concentrations. mCherry-OLS1 was used at 5 μM. Scale bar: 10μm.

(C) *In vitro* PS of OLS1 at 50mM NaCl. OLS1 was used at 5 μM. Different concentration of mCherry-ILS1 were added to the mix. Scale bar: 10μm.

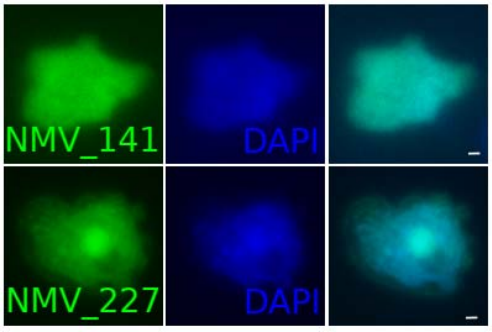
(D) Immunofluorescence demonstrating localization of client proteins from the OL of the VF. Proteins were endogenously tagged with 3xHA at the C-terminal and infection was carried out for 6 hours before fixation. VFs were labelled using DAPI. Scale bar: 1μm.

(E) Schematic representation of the vector and knock-in (KI) strategy utilized for endogenous tagging of *r322* and *r336/337*. Selection cassette was introduced by homologous recombination and recombinant viruses were generated, selected and cloned. *nat*: Nourseothricin N-acetyl transferase. Primers annealing locations are shown and successful KI as clonality is demonstrated by PCR. Expected sizes are indicated in the figure.

A


distance
rbf
linear
rbf
linear
rbf
linear
rbf
linear
ParseV2
Molphase
SVC classifiers
Existing methods
Fraction of shared positive predictions
1.00
0.75
0.50
0.25
Number of metrics
2
5
15
20
23
Existing methods
SVC classifiers
NMV_095
NMV_141
NMV_227
NMV_238

B

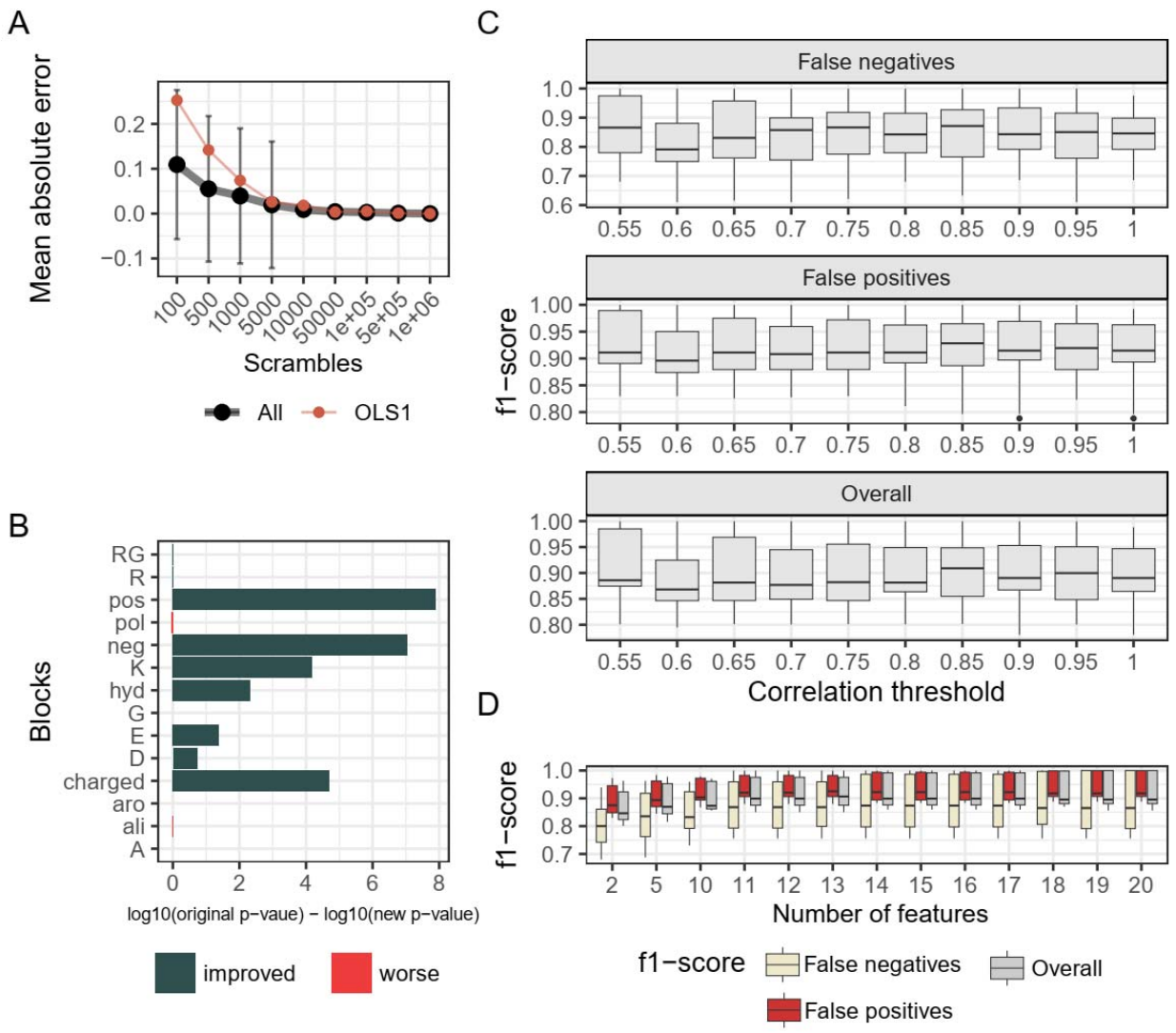

NMV_141
DAPI
NMV_227
DAPI

1069

Supplementary Figure 4. Exploration of different methods to predict scaffold proteins in noumeavirus with OLS1 and homologs.

(A) Correlation matrix between different classifiers positive proteins predictions. White number in each cell indicates the number of shared predictions between two classifiers. Numbers in black give the number of representative genomes in which the two classifiers predicted at least one scaffold protein. The cell color corresponds to the number of shared predictions normalized by the number of positive proteins in the classifier that identified the lowest number of positive proteins. Noumeavirus 4 proteins predicted as positive by at least one of our methods are shown in the graph under the heatmap, in grey, red, orange, and blue respectively. Each line corresponds to a link between the protein and associated classifier.

(B) *A. castellanii* cells expressing C-terminally tagged NMV_141 or NMV_227. Scale bar: 1µm.



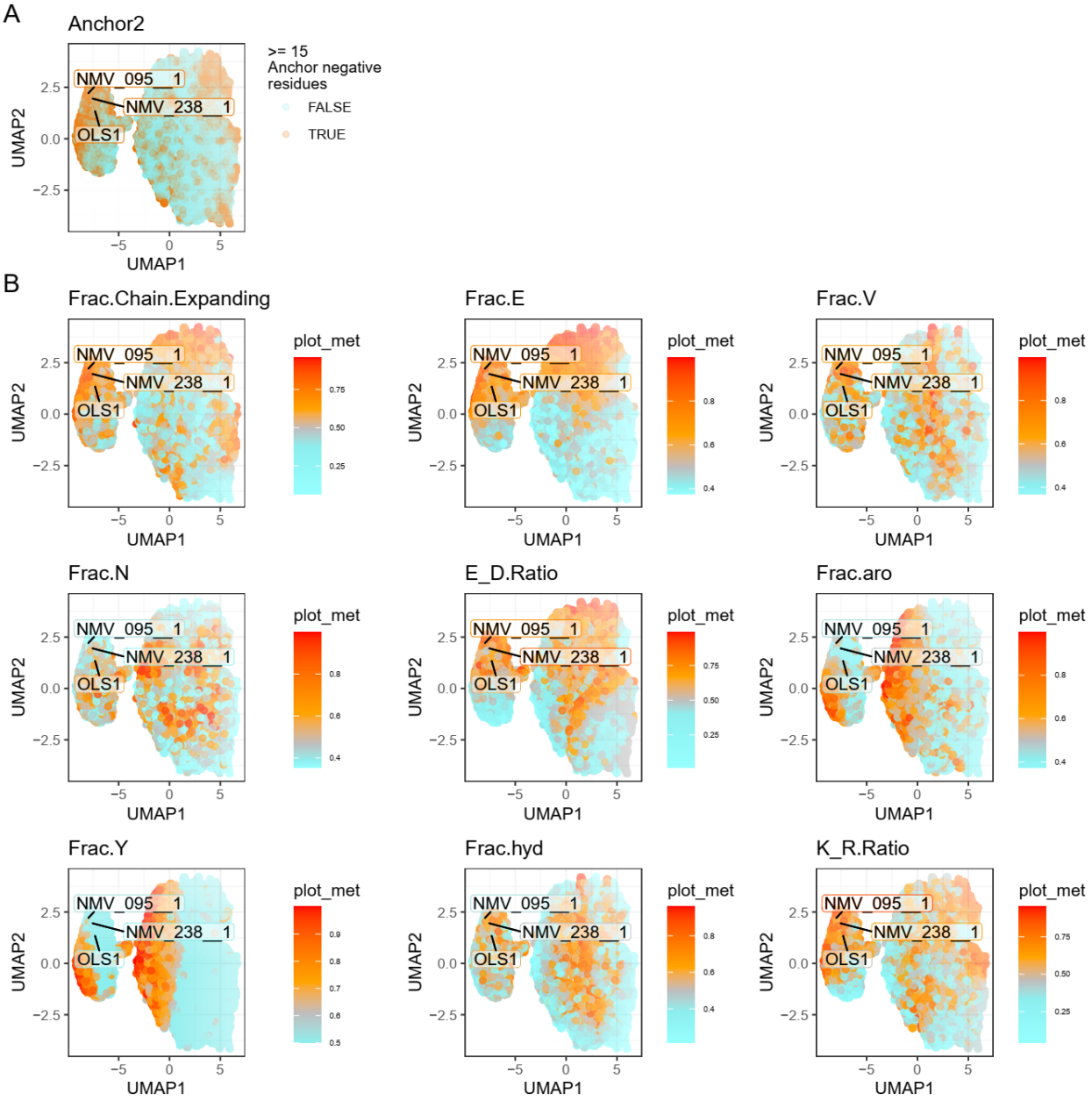
Supplementary Figure 5. Improvement of feature calculations and optimization of classifiers

(A) Optimal number of bootstraps performed by Nardini to estimate a Z-score. The error is given as the absolute difference with the Z-score obtained using 1 million sequence randomizations of mimivirus IDRs. Only Nardini features with an inter-quantile range above 0 are considered.

(B) Modification of the block calculation method from King *et al.*[22]. The wilcoxon p-values for the reference IDR and homologs against the rest of mimivirus IDRs were compared. Note: To ensure that the features were shared and relevant, we only kept the ones with significantly lower variance in the predicted scaffold proteins compared to the negative class.

(C) F1-score of classifiers with OLS1, NMV_238 and NMV_095 and homologs as training set compared to the correlation threshold above which features can be clustered together. The final classifier clusters features with a spearman correlation coefficient above 0.55, as correlation thresholds of 0.55 and higher all provide high f1-scores.

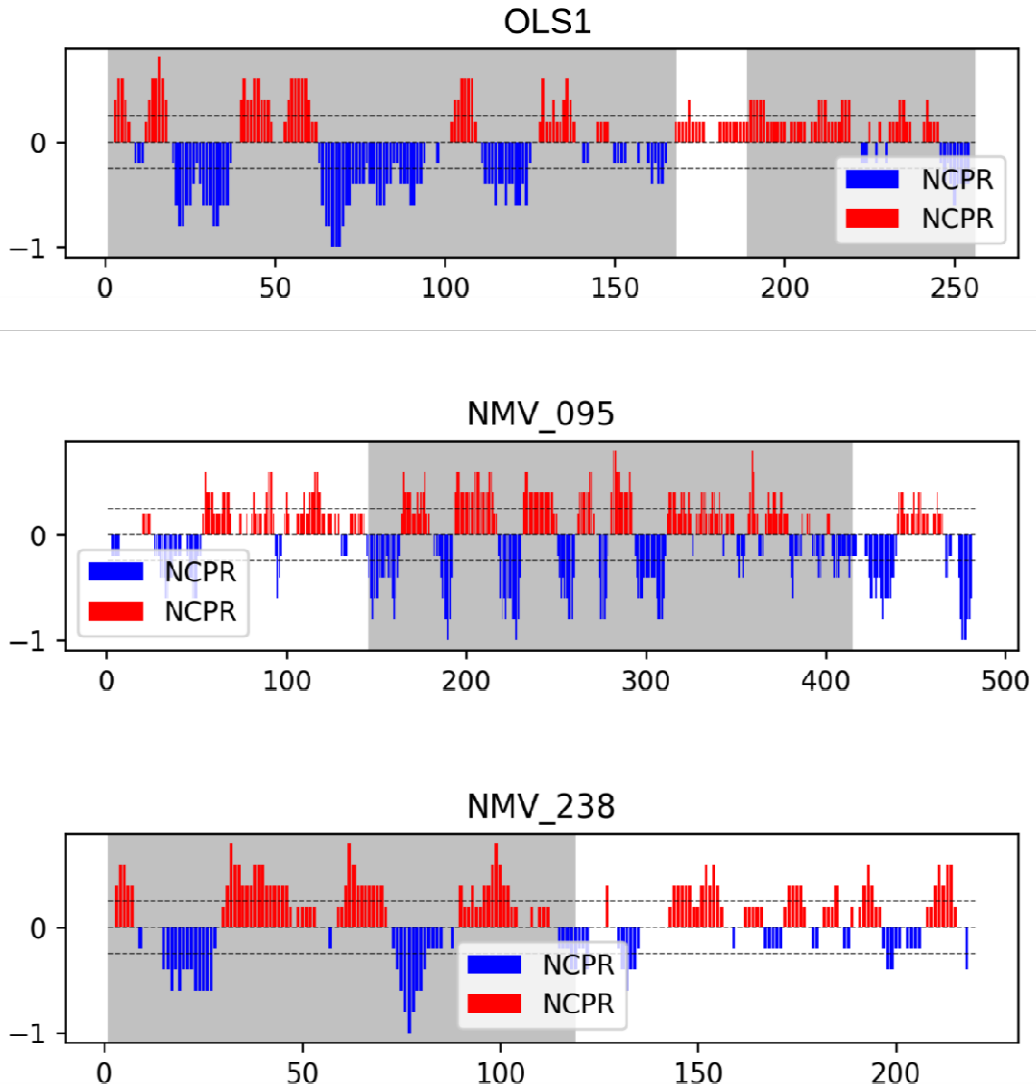
(D) F1-score variations according to the number of features considered by the classifier. The final classifier is based on selected 11 uncorrelated features, as the f1-scores do not improve when using more uncorrelated features.



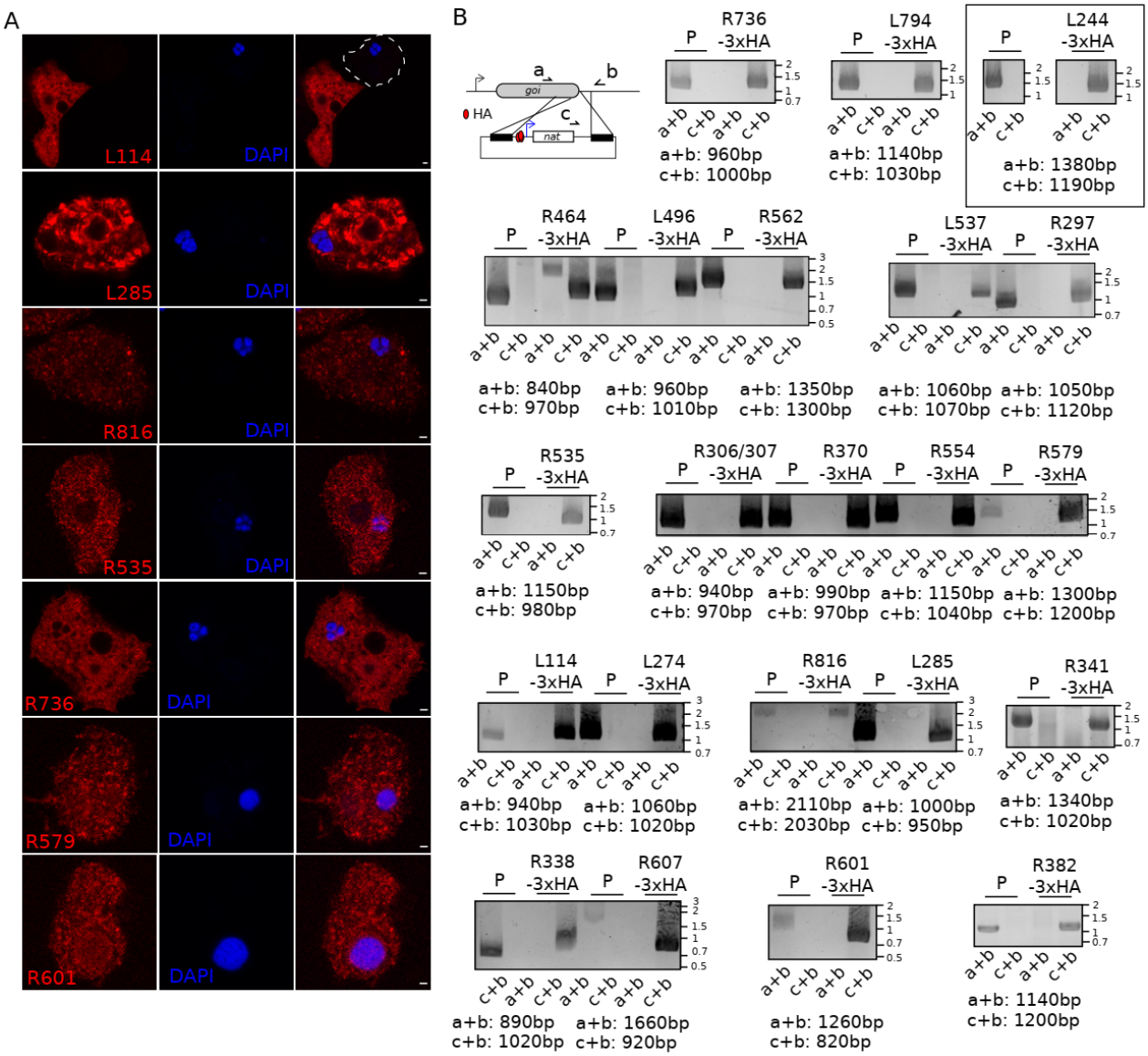
Supplementary Figure 6. Distribution of classifier features and control features across the UMAP representation of IDRs of *Nucleocytoviricota*.

(A) Anchor2 values across the UMAP. Red points are for possible truly disordered IDRs having at least 15 amino-acids with a Anchor2 score under 0.5.

(B) Final classifier features across the UMAP representation of IDRs based on those 19 features plus the positive-positive and positive-negative Nardini Z-score shown in Figure 4, highlighting each feature contribution to the classifier and in validated scaffold proteins. Frac. goes for "fraction" and the chain expanding residues are E, D, R, K and P. PPII is the propensity to form polyproline II conformations. "pol" goes for polar residues and "hyd" for hydrophobic.



Supplementary Figure 7. Charge segregation in reference scaffold proteins. The positive and negative charge distribution given by CIDER is shown as red for positive and blue for negative. The position of the IDRs is highlighted by the grey-shaded rectangle. These 3 identified scaffold proteins show similar alternating positive and negative patches across the IDR. NCPR: Net charge per residue.

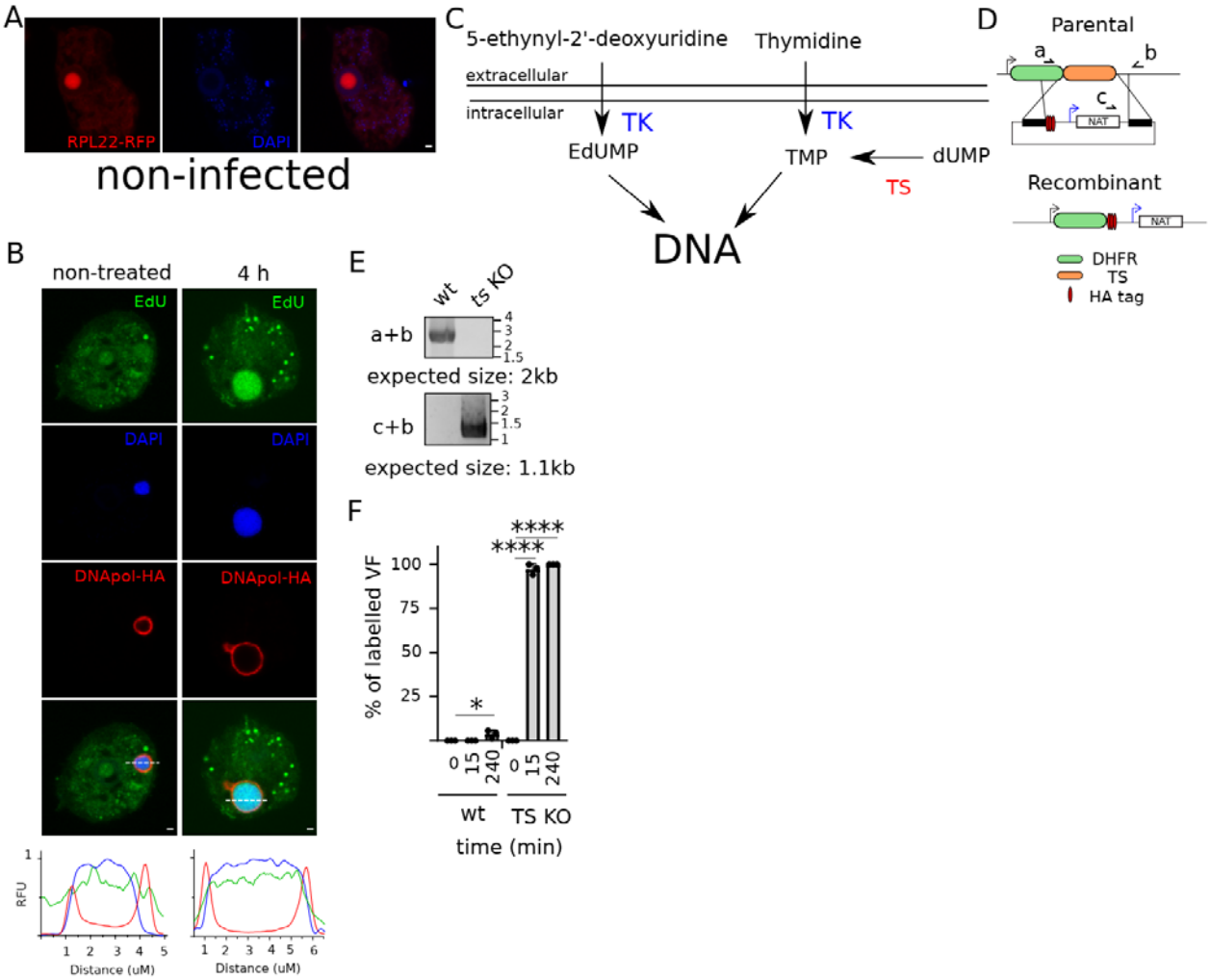


Supplementary Figure 8.

(A) Immunofluorescence demonstrating localization of proteins identified at the VF in [11] that could not be confirmed by endogenous tagging. Proteins were endogenously tagged with

3xHA at the C-terminal and infection was carried out for 6 hours before fixation. VFs were labelled using DAPI. Scale bar: 2µm.

(B) Schematic representation of the vector and knock-in (KI) strategy utilized for endogenous tagging of client proteins. Selection cassette was introduced by homologous recombination and recombinant viruses were generated, selected and cloned. *nat*: Nourseothricin N-acetyl transferase. Primers annealing locations are shown and successful KI as clonality is demonstrated by PCR. Expected sizes are indicated in the figure.



Supplementary Figure 9.

(A) *A. castellanii* cells expressing RPL22-RFP. RPL22-RFP was detected in the nucleolus and the cytoplasm of non-infected cells. DAPI: DNA. Scale bar: 2 µm.

(B) Detection of DNA synthesis by EdU labelling. Viral infection by wild type viruses was allowed to proceed for 4-6 hours and labelling time is indicated. IL of the VFs was labelled using DAPI. Scale bar: 2 µm.

(C) Cartoon representing the two pathways for incorporation of deoxy-thymidine into the DNA. TK: Thymidylate kinase. TS: Thymidylate Synthase.

(D) Cartoon representing the strategy to disrupt the Thymidylate Synthase gene without disrupting the N-terminal Dihydrofolate reductase domain.

(E) Efficient disruption of *thymidylate synthase* of mimivirus and clonality of recombinant viruses was demonstrated by PCR. Primers annealing locations are shown in Figure S6C. Expected sizes are indicated in the figure.

(F) Quantification of the relative number of VFs with EdU labeling in wild type viruses or *thymidylate synthase* KO is shown. Data correspond to the mean ± SD of 3 independent experiments. At least 100 VFs were counted during each experiment. ns ($P > 0.05$), * ($P \leq 0.05$), ** ($P \leq 0.01$), *** ($P \leq 0.001$) and **** ($P \leq 0.0001$).

Supplementary Table 1. MS-based characterization of R505, R562 and R336-R337 interactomes.

Supplementary Table 2. Prediction of the IDRome and putative scaffold proteins throughout the *Nucleocytoviricota*.

Supplementary Table 3. MS-based quantitative proteomic identification of OLS1(R561)-GFP and ILS1(R252)-RFP binding partners.

Supplementary Table 4. Primers used in this study.