

UNIVERSIDADE FEDERAL DE MINAS GERAIS
INSTITUTO DE CIÊNCIAS BIOLÓGICAS
PROGRAMA DE PÓS-GRADUAÇÃO EM MICROBIOLOGIA

Amanda Stéphanie Arantes Witt

**ANÁLISE GLOBAL E GENE A GENE DO GC% DE VÍRUS GRANDES E GIGANTES
PERTENCENTES AO FILO *NUCLEOCYTOVIRICOTA***

Belo Horizonte

2022

Amanda Stéphanie Arantes Witt

**ANÁLISE GLOBAL E GENE A GENE DO GC% DE VÍRUS GRANDES E GIGANTES
PERTENCENTES AO FILO *NUCLEOCYTOVIRICOTA***

Versão final

Dissertação de Mestrado apresentada ao Programa de Pós-Graduação em Microbiologia do Instituto de Ciências Biológicas da Universidade Federal de Minas Gerais, como requisito parcial para a obtenção do Grau de Mestre em Microbiologia.

Orientador: Prof. Jônatas Santos Abrahão

Belo Horizonte

2022

043

Witt, Amanda Stéphanie Arantes.

Análise global e gene a gene do GC% de vírus grandes e gigantes pertencentes ao filo Nucleocytoviricota [manuscrito] / Amanda Stéphanie Arantes Witt. – 2022.

94 f. : il. ; 29,5 cm.

Orientador: Prof. Jônatas Santos Abrahão.

Dissertação (mestrado) – Universidade Federal de Minas Gerais, Instituto de Ciências Biológicas. Programa de Pós-Graduação em Microbiologia.

1. Microbiologia. 2. Genômica. 3. Vírus gigantes. 4. Composição de Bases. 5. Poxvírus. I. Abrahão, Jônatas Santos. II. Universidade Federal de Minas Gerais. Instituto de Ciências Biológicas. III. Título.

CDU: 579



UNIVERSIDADE FEDERAL DE MINAS GERAIS
INSTITUTO DE CIÊNCIAS BIOLÓGICAS
CURSO DE PÓS-GRADUAÇÃO EM MICROBIOLOGIA

ATA DE DEFESA DE DISSERTAÇÃO

ATA DA DEFESA DE DISSERTAÇÃO DE **AMANDA STÉPHANIE ARANTES WITT**

Nº REGISTRO: **2020682634**

Às 14:00 horas do dia **02 de MAIO de 2022**, reuniu-se, por via remota, a Comissão Examinadora composta pelos Drs. Giliane de Souza Trindade (Departamento de Microbiologia/ICB/UFMG), Juliana Reis Cortines (Universidade Federal do Rio de Janeiro) e o Prof. Dr. Jônatas Santos Abrahão – Orientador, para julgar o trabalho final "**Análise global e gene a gene do GC% de vírus grandes e gigantes pertencentes ao filo *Nucleocytoviricota***", da aluna **Amanda Stéphanie Arantes Witt**, requisito final para a obtenção do Grau de **MESTRA EM CIÊNCIAS BIOLÓGICAS: MICROBIOLOGIA**. Abrindo a sessão, o Presidente da Comissão, Prof. Jônatas Santos Abrahão - orientador, após dar a conhecer aos presentes o teor das Normas Regulamentares do Trabalho Final, passou a palavra à candidata, para a apresentação de seu trabalho. Seguiu-se a arguição pelos Examinadores, com a respectiva defesa da candidata. Logo após, a Comissão se reuniu, sem a presença da candidata e do público, para julgamento e expedição de resultado final. A candidata foi considerada **APROVADA**. O resultado final foi comunicado publicamente à candidata pelo Presidente da Comissão. Nada mais havendo a tratar, o Presidente encerrou a reunião e lavrou a presente ata, que será assinada por todos os membros participantes da Comissão Examinadora. O candidato tem 60 (sessenta) dias, a partir desta data, para entregar a versão final da dissertação ao Programa de Pós-graduação em Microbiologia da UFMG e requerer seu diploma.

Membros da banca:

Profa. Dra. Giliane de Souza Trindade

Dra. Juliana Reis Cortines

De acordo:

Prof. Dr. Jônatas Santos Abrahão
(orientador)

Profa. Danielle da Glória de Souza
(Coordenadora do PPGM/UFMG)

Belo Horizonte, 02 de maio de 2021



Documento assinado eletronicamente por **Jonatas Santos Abrahao, Professor do Magistério Superior**, em 03/05/2022, às 09:49, conforme horário oficial de Brasília, com fundamento no art. 5º do [Decreto nº 10.543, de 13 de novembro de 2020](#).



Documento assinado eletronicamente por **Daniele da Gloria de Souza, Coordenador(a) de curso de pós-graduação**, em 03/05/2022, às 16:05, conforme horário oficial de Brasília, com fundamento no art. 5º do [Decreto nº 10.543, de 13 de novembro de 2020](#).



Documento assinado eletronicamente por **Giliane de Souza Trindade, Chefe**, em 04/05/2022, às 11:46, conforme horário oficial de Brasília, com fundamento no art. 5º do [Decreto nº 10.543, de 13 de novembro de 2020](#).



Documento assinado eletronicamente por **Juliana Reis Cortines, Usuário Externo**, em 04/05/2022, às 12:59, conforme horário oficial de Brasília, com fundamento no art. 5º do [Decreto nº 10.543, de 13 de novembro de 2020](#).



A autenticidade deste documento pode ser conferida no site https://sei.ufmg.br/sei/controlador_externo.php?acao=documento_conferir&id_orgao_acesso_externo=0, informando o código verificador **1363398** e o código CRC **47156B40**.

AGRADECIMENTOS

Agradeço a todos os meus queridos professores, que puderam contribuir com tanta riqueza para minha formação profissional e pessoal, e aos meus colegas de laboratório, que sempre incentivaram e colaboraram de diferentes formas para o desenvolvimento deste trabalho.

Sou grata, em especial, ao meu orientador e professor, Jônatas Abrahão, pelo acolhimento, pela paciência nos momentos mais difíceis, pelo incentivo e suporte essenciais neste momento tão delicado de pandemia da COVID-19, pelo carinho e atenção não apenas comigo, mas para com cada um de seus alunos, e principalmente, por confiar em mim e me conceder a oportunidade de realizar este trabalho.

Agradeço ao grupo de pesquisa GEPVIG, ao Laboratório de Vírus, ao Instituto de Ciências Biológicas e à Universidade Federal de Minas Gerais, pelo espaço disponibilizado e por todo o suporte e incentivo oferecido para a realização deste trabalho.

Agradeço também às agências de fomento, CNPq, Capes e Fapemig, pela viabilização e importante apoio financeiro ao projeto de pesquisa.

E a todos os demais aqui não mencionados, mas que me deram tanto apoio, carinho e doses diárias de esperança, para que eu chegasse até aqui e pudesse, enfim, realizar este sonho – meus pais, minha irmã, minha família, meus amigos, minhas médicas, e a todos os que torcem por mim, minha mais sincera e profunda gratidão.

Muito obrigada!

Resumo

O filo *Nucleocytoviricota* (ou NCLDV) é um dos grupos taxonômicos mais diversos, incluindo representantes com os maiores genomas observados na virosfera conhecida. Uma das métricas de estudo genômico em composição de sequências é a proporção percentual de guaninas e citosinas, também conhecida por porcentagem de conteúdo GC, que traz importantes informações acerca da biologia dos organismos. No presente trabalho, o GC% de 61 genomas de vírus pertencentes a *Nucleocytoviricota* foi analisado, gene a gene. Dentre estes, estão representantes de *Phycodnaviridae*, *Mimiviridae*, *Iridoviridae*, *Marseilleviridae*, *Asfarviridae*, *Poxviridae* e “*Pithoviridae*”, bem como linhagens ainda não classificadas. Os genomas de referência foram obtidos a partir do banco de dados genômicos público GenBank (NCBI), e foram criados arquivos *multifasta* contendo somente as sequências codificantes anotadas para cada genoma. Estes arquivos foram submetidos ao programa de livre acesso *InfoSeq* do *EMBOSS* (EMBL-EBI) e assim obteve-se a porcentagem de conteúdo GC de cada CDS, para cada genoma. De porte destes dados, foram avaliadas as distribuições e variações de GC% em relação a CDS ao longo dos genomas individualmente, dentro de cada grupo taxonômico e globalmente em *Nucleocytoviricota*. Nossos resultados revelaram como NCLDV é um filo diverso, não somente na biologia viral de seus integrantes, mas também na composição dos seus genomas. A amplitude dos valores de GC% observados em *Nucleocytoviricota* é significativa, e membros de algumas famílias, como *Mimiviridae* e *Poxviridae*, ganham destaque. Embora a maioria dos genes analisados apresentem GC% próximo à média das CDS, foi observado um importante número de genes com GC% notavelmente elevado ou abaixo da média. Também foram observados, em diferentes genomas, agrupamentos de genes em regiões variáveis, cujos valores percentuais de GC eram significativamente mais elevados que a média. Foram selecionados, posteriormente, genes considerados *outliers*, de valores máximos e mínimos observados em cada genoma, afim de caracterizar as CDS de valores extremos e verificar possíveis tendências evolutivas dos grupos ou conservação de genes ortólogos. Este trabalho trouxe, pela primeira vez, uma análise global e gene-a-gene da distribuição e o perfil da variação percentual GC nos genomas de *Nucleocytoviricota*,

ilustrando uma parte da diversidade deste grupo e identificando novos alvos para estudos futuros.

Palavras-chave: Conteúdo GC%. Genômica. *Nucleocytoviricota*. Poxvírus; Vírus gigantes.

Abstract

Nucleocytoviricota (or NCLDV) is one of the most diverse taxonomic groups, including representatives with the largest genomes observed in the known virosphere. Genomic studies on sequence composition counts with many different approaches, one of which calculates the proportion of guanines and cytosines within the sequence, known as GC content. The GC% content can lead to important information about the organism's biology and various biological aspects. In the present work, the GC% of 61 genomes belonging to *Nucleocytoviricota* was calculated gene by gene. Among these genomes are the representatives of *Phycodnaviridae*, *Mimiviridae*, *Iridoviridae*, *Marseilleviridae*, *Asfaridae*, *Poxviridae* and "Pithoviridae", as well as unclassified lineages. Reference genomes were obtained from the public genomic database GenBank (NCBI), and multifasta files were created containing only coding sequences annotated for each genome. These files were submitted to the sequence content analysis and free access program InfoSeq, by EMBOSS (EMBL-EBI). We then obtained the GC% content of each CDS, for each genome individually, within each group and globally in *Nucleocytoviricota*. Our results revealed how NCLDV is diverse, not only considering the viral biology of its members, but also in the GC% composition of its genomes. The range of GC% values observed in *Nucleocytoviricota* is significant, and some families, such as *Mimiviridae* and *Poxviridae*, gain prominence. Although most of the analyzed genes presented GC% values close to the CDS mean, an important number of genes with GC% significantly over or below the mean was observed. Furthermore, different genomes presented clusters of genes in defined regions with percentage values of GC significantly above the average. Outlier gene sequences of each group were evaluated in order to determine presence of orthologs or potential evolutionary tendencies shared among these groups. This work brings, for the first time, a global and gene-by-gene analysis of the distribution and profile of the GC content percentage within genomes, illustrating part of the diversity within *Nucleocytoviricota*, and identifying possible new targets for future studies.

Key-words: GC% content. Genomics. *Nucleocytoviricota*. Poxviruses; Giant viruses.

LISTA DE ILUSTRAÇÕES

Figura 1 – Vírus do mosaico do tabaco em microscopia eletrônica	15
Figura 2 – <i>Acanthamoeba polyphaga mimivirus</i> (APMV), inicialmente denominado Coco de Bradford, no interior de uma ameba de vida livre	17
Figura 3 – Mimivirus em células infectadas e em sobrenadante livre de células	18
Figura 4 – Tupanvírus soda lake, partículas e ciclo	19
Figura 5 – Fluxograma de esquema do trabalho realizado	25
Figura 6 – Representação da dispersão dos valores de conteúdo GC%, gene a gene, em genomas de <i>Phycodnaviridae</i>	34
Figura 7 – Representação da dispersão dos valores de conteúdo GC%, gene a gene, em genomas de <i>Mimiviridae</i> (painel A).	36
Figura 8 – Representação da dispersão dos valores de conteúdo GC%, gene a gene, em genomas de <i>Mimiviridae</i> (painel B).	38
Figura 9 – Representação da dispersão dos valores de conteúdo GC%, gene a gene, em genomas de <i>Iridoviridae</i>	40
Figura 10 – Representação da dispersão dos valores de conteúdo GC%, gene a gene, em genomas de <i>Marseilleviridae</i> .	41
Figura 11 – Representação da dispersão dos valores de conteúdo GC%, gene a gene, em genomas de <i>Asfarviridae</i> .	43
Figura 12 – Representação da dispersão dos valores de conteúdo GC%, gene a gene, em genomas de <i>Poxviridae</i> (painel A).	45
Figura 13 – Representação da dispersão dos valores de conteúdo GC%, gene a gene, em genomas de <i>Poxviridae</i> (painel B).	46

Figura 14 – Representação da dispersão dos valores de conteúdo GC%, gene a gene, em genomas de *Entomopoxvirinae*. 47

Figura 15 – Representação da dispersão dos valores de conteúdo GC%, gene a gene, em genomas de "Pithoviridae". 49

Figura 16 – Mapas de classificação do valor de conteúdo GC% por cores, gene a gene, individualmente para cada genoma e intra-grupos taxonômicos. - *Phycodnaviridae*. 52

Figura 17 – Mapas de classificação do valor de conteúdo GC% por cores, gene a gene, individualmente para cada genoma e intra-grupos taxonômicos. - *Mimiviridae*. 54

Figura 18 – Mapas de classificação do valor de conteúdo GC% por cores, gene a gene, individualmente para cada genoma e intra-grupos taxonômicos. - *Iridoviridae*. 56

Figura 19 – Mapas de classificação do valor de conteúdo GC% por cores, gene a gene, individualmente para cada genoma e intra-grupos taxonômicos. - *Marseilleviridae* 58

Figura 20 – Mapas de classificação do valor de conteúdo GC% por cores, gene a gene, individualmente para cada genoma e intra-grupos taxonômicos. - *Asfarviridae*. 60

Figura 21 – Mapas de classificação do valor de conteúdo GC% por cores, gene a gene, individualmente para cada genoma e intra-grupos taxonômicos. - *Poxviridae* 63

Figura 22 – Mapas de classificação do valor de conteúdo GC% por cores, gene a gene, individualmente para cada genoma e intra-grupos taxonômicos. - *Pithoviridae*. 65

Figura 23 – Mapas de classificação do valor de conteúdo GC% por cores, gene a gene, individualmente para cada genoma e intra-grupos taxonômicos. - *Nucleocytoviricota*. 68

LISTA DE TABELAS

TABELA 1 – Total de 61 genomas virais pertencentes ao filo <i>Nucleocytoviricota</i> selecionados e utilizados neste trabalho	33
TABELA 2 – Valores GC% máximos e mínimos observados em <i>Phycodnaviridae</i>	70
TABELA 3 – Valores GC% máximos e mínimos observados em <i>Mimiviridae</i>	72
TABELA 4 – Valores GC% máximos e mínimos observados em <i>Iridoviridae</i>	73
TABELA 5 – Valores GC% máximos e mínimos observados em <i>Marseilleviridae</i>	74
TABELA 6 – Valores GC% máximos e mínimos observados em <i>Asfarviridae</i>	75
TABELA 7 – Tabela A de Poxviridae. Valores GC% máximos e mínimos observados em <i>Poxviridae</i>	77
TABELA 8 – Tabela B de Poxviridae. Valores GC% máximos e mínimos observados em <i>Poxviridae</i>	78
TABELA 9 – Tabela C de Poxviridae. Valores GC% máximos e mínimos observados em <i>Poxviridae</i>	79
TABELA 10 – Valores GC% máximos e mínimos observados em “Pithoviridae”	80

LISTA DE SIGLAS E ABREVIATURAS

Kb	Quilobase
DNA	Ácido desoxirribonucleico
RNA	Ácido ribonucleico
Nm	Nanômetro
µm	Micrometro
TEM	Microscopia eletrônica de transmissão
MEV	Microscopia eletrônica de varredura
CDS	<i>“coding DNA sequence”</i> ou sequência de DNA codificadora
GC	Guanina+Citosina
%	Porcentagem
IL-1	Interleucina-1

Sumário

1. INTRODUÇÃO	15
1.1 A virosfera e sua diversidade	15
1.1.1 A descoberta dos vírus	15
1.1.2 O início dos estudos sobre a virosfera	16
1.1.3 A virosfera que conhecemos hoje	17
1.2 O estudo de sequências	20
1.3 Estudos de composição de sequências biológicas e o conteúdo G+C	21
1. JUSTIFICATIVA	23
2. OBJETIVOS	24
3.1 Objetivo geral	24
3.2 Objetivos específicos	24
3. FLUXOGRAMA	25
4. MATERIAL E MÉTODO	26
5.1 Seleção de genomas e aquisição de sequências	26
5.2 Processamento de genomas e determinação de conteúdo GC%	27
5.3 Avaliação de GC% gene a gene	28
5.3.1 Construção de painéis de dispersão	28
5.3.2 Construção dos mapas de variação de GC%	29
5. RESULTADOS	31
6.1 Banco de dados e genomas utilizados	31
6.2 Variação do perfil de conteúdo GC%	33
6.3 Comparação da variação de GC% gene a gene intra-grupos	50
6.4 Regiões codificadoras de percentagens GC% máximas e mínimas	70
6. DISCUSSÃO	81
7.1 Extração de informações biológicas por análises <i>in silico</i> e a obtenção do percentual de GC de regiões específicas	81
7.2 Variação global de GC% em <i>Nucleocyotoviricota</i>	82
7.3 Variação de GC% dentro das famílias virais	83
7.4 Alvos para trabalhos futuros	86
7. CONCLUSÃO	88

REFERÊNCIAS BIBLIOGRÁFICAS.....	89
--	-----------

PRODUÇÕES CIENTÍFICAS	93
------------------------------------	-----------

I. Artigos aceitos para publicação	93
--	----

II. Resumo publicado em evento científico	93
---	----

1. INTRODUÇÃO

1.1 A virosfera e sua diversidade

1.1.1 A descoberta dos vírus

A virologia, ao contrário de outras grandes áreas da biologia, possui uma história ainda muito recente, com cerca de 130 anos desde a descoberta do primeiro agente viral em 1892, pelos pesquisadores Adolf Mayer, Dimitri Ivanovsky e Martinus Beijerinck, que estudavam a doença do mosaico do tabaco. Em seus trabalhos, Beijerinck, inicialmente, relatou ter observado algo que descreveu como “*contagium vivum fluidum*”, que ia de encontro com o que se conhecia a respeito de agentes causadores de doenças, que obrigatoriamente deveriam ser organismos celulares. A partir destes e de outros estudos, o primeiro agente viral foi isolado e denominado vírus do mosaico do tabaco (BOS, 1888; LECOQ, 2001; LUSTIG; LEVINE, 1992).



Figura 1. Vírus do mosaico do tabaco em microscopia eletrônica. Fonte: (KLUG, 1999).

O termo “vírus”, hoje mundialmente difundido, foi, então, cunhado e, posteriormente, associado ao conceito popular de agentes filtráveis, não visíveis por microscopia óptica e, no entanto, capazes de causar doença em células, tecidos e organismos. Pouco tempo depois, foram isolados o primeiro vírus animal, vírus da febre aftosa (1898), e o primeiro vírus capaz de infectar humanos, o vírus da febre amarela (1901) (BOS, 1888; LECOQ, 2001; LUSTIG; LEVINE, 1992).

1.1.2 O início dos estudos sobre a virosfera

Com a difusão e popularização do termo vírus, fortaleceram-se ideias acerca do que estes agentes biológicos seriam e, desta forma, surgiram conceitos que definiriam, por anos, a identidade dos vírus. Alguns destes conceitos preconizavam que todo organismo, para ser considerado um agente viral, deveria ser completamente dependente da maquinaria celular de seu organismo hospedeiro, portador de um genoma naturalmente reduzido e codificante para poucas proteínas. A partir daí, os vírus passaram a ser descritos como parasitas ubíquos e capazes de infectar organismos dos três domínios da vida (LWOFF, 1957).

Naturalmente, os vírus se popularizaram como agentes causadores das mais diversas doenças – tais como a febre amarela, influenza, hepatite, sarampo, varíola, raiva, etc. – tanto em humanos quanto em outros animais, domésticos e não domésticos (BELLINI; ROTA; ROTA, [S.d.]; BOUVIER; PALESE, 2008; FISHER; STREICKER; SCHNELL, 2018; MONATH, 2001; PURCELL, 1994). Também, por causa da grande variedade de doenças atribuídas aos vírus, nós pudemos isolar e caracterizar uma parcela considerável dos agentes virais conhecidos atualmente – o que, inquestionavelmente, condicionou a virosfera conhecida a um viés de estudo médico e primariamente de interesse humano.

No entanto, hoje podemos estimar, em algum nível, a real imensidão e diversidade que se encontra na virosfera. Portando deste conhecimento é que sabemos que estes microrganismos fantásticos, na verdade, estão em sua maioria muito distantes de ter relação com alguma doença ou similar. Além disto, o papel negativo associado aos vírus

construído e alimentado por anos, é de se considerar equivocado e injusto, pois estes são importantes agentes em diversos nichos ecológicos, desempenhando funções essenciais aos ecossistemas onde habitam (KIMURA *et al.*, 2008; WEINBAUER, 2004; WILHELM; SUTTLE, [S.d.]; ZHANG; GUI, 2018).

1.1.3 A virosfera que conhecemos hoje

A ideia associada ao que era um vírus e seu nicho começou a mudar em 1992, quando o primeiro vírus gigante de amebas foi isolado por Tim Rowbothan e colaboradores do Laboratório de Saúde Pública de Leeds (Inglaterra). O isolado, proveniente de material coletado para investigação acerca de um surto de pneumonia em um hospital da cidade de Bradford, inicialmente foi identificado como uma bactéria e batizado “coco de Bradford” (RAOULT; LA SCOLA; BIRTLES, 2007) .

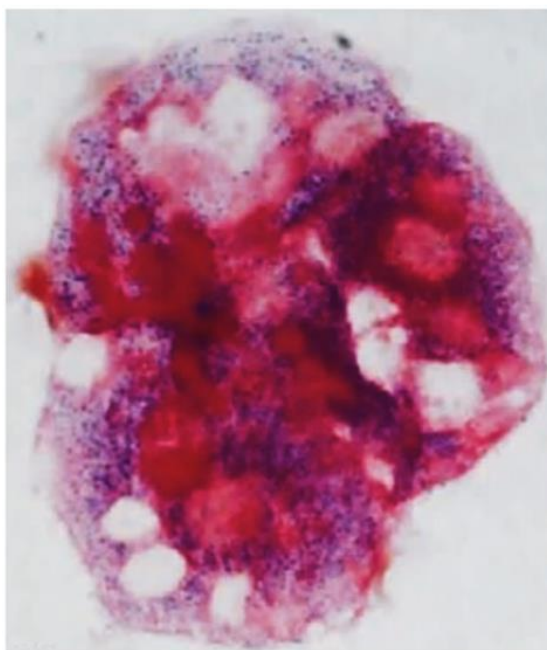


Figura 2. Acanthamoeba polyphaga mimivirus (APMV), inicialmente denominado Coco de Bradford, no interior de uma ameba de vida livre. APMV corado de violeta através da coloração de Gram, visível por microscopia óptica. Fonte: (RAOULT; LA SCOLA; BIRTLES, 2007).

Até então, corado pela coloração de Gram no interior de amebas do gênero *Acanthamoeba*, estes organismos não pareciam tão diferentes de outras bactérias já conhecidas. Apesar disso, durante o processo de cultivo e identificação molecular utilizando iniciadores bacterianos universais para a região 16S e para DNA ribossomal, notou-se que havia algo de diferente com aqueles microrganismos, já que não se obtinham resultados positivos durante estes procedimentos rotineiros e bem consolidados na identificação bacteriana.

Após anos de interrupção dos estudos com os misteriosos organismos de Bradford, o grupo de pesquisa coordenado por Bernard La Scola e Didier Raoult, da Aix Marseille Université, França, conseguiu efetivamente identificar o primeiro vírus gigante infectando amebas, o *Acanthamoeba polyphaga* mimivirus (APMV), em 2003 (GREUB; RAOULT, 2004; SCOLA et al., 2003).

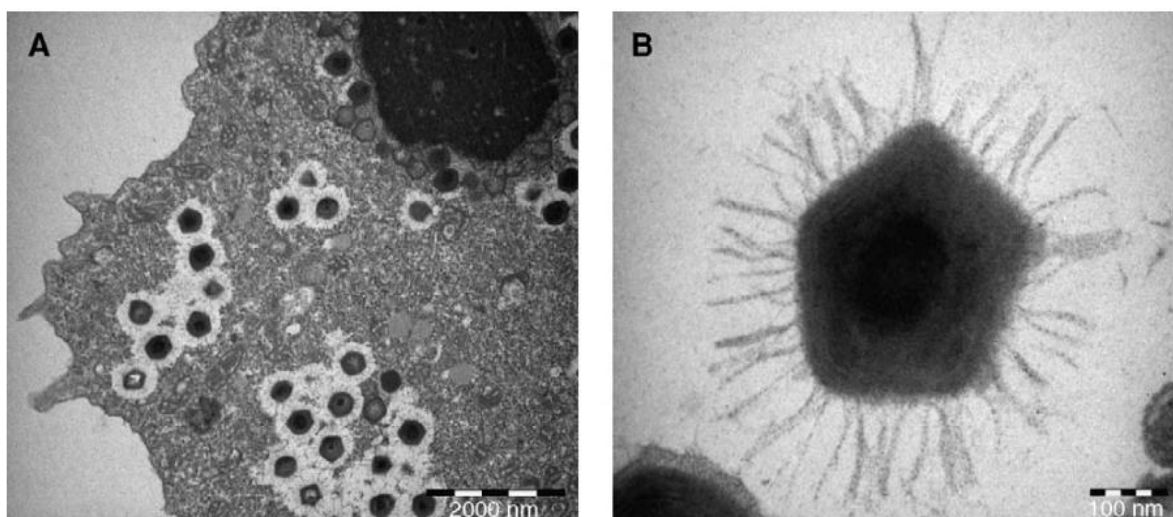


Figura 3. Mimivirus em células infectadas e em sobrenadante livre de células. Microscopia eletrônica de transmissão de *A. polyphaga* infectada por Mimivírus 8h p.i. mostra o acúmulo intracitoplasmático de partículas virais (A). Vírion icosaédrico não envelopado circundado por fibrilas (B). Fonte: (SUZAN-MONTI; LA SCOLA; RAOULT, 2006).

Desde sua descoberta, novas portas se abriram no universo da virosfera conhecida até aquele momento, o que incluiu o desenvolvimento e validação de novas técnicas de isolamento e identificação de agentes virais. Desta forma, diversos novos vírus gigantes foram sendo descobertos nos anos subsequentes, ao redor de todo o mundo. Ainda, com a descrição destes vírus, a discussão acerca do conceito de vírus tornou-se mais complexa, já que estes gigantes microrganismos trouxeram consigo gigantes perguntas (FORTERRE, 2010).

Os vírus estão, hipoteticamente, em todos os ambientes em que organismos celulares estão e sob as mais diversas pressões ambientais. Alguns dos vírus gigantes descobertos em lugares considerados mais extremos são os tupanvírus, isolados de lagoas salinas no Pantanal e de amostras oceânicas extraídas de extensa profundidade na costa brasileira, os pithovirus e mollivirus, isolados a partir de amostras de mais de 30mil anos de idade congeladas e extraídas do *permafrost* na Sibéria, mimivírus isolado a partir de águas de baixa temperatura na Antártida e medusavírus, isolado a partir de amostras de água de fontes termais no Japão, por exemplo (ABERGEL; CLAVERIE, 2014; ABRAHÃO *et al.*, 2018b; ANDRADE *et al.*, 2018; LEGENDRE *et al.*, 2015; YOSHIKAWA *et al.*, 2019)

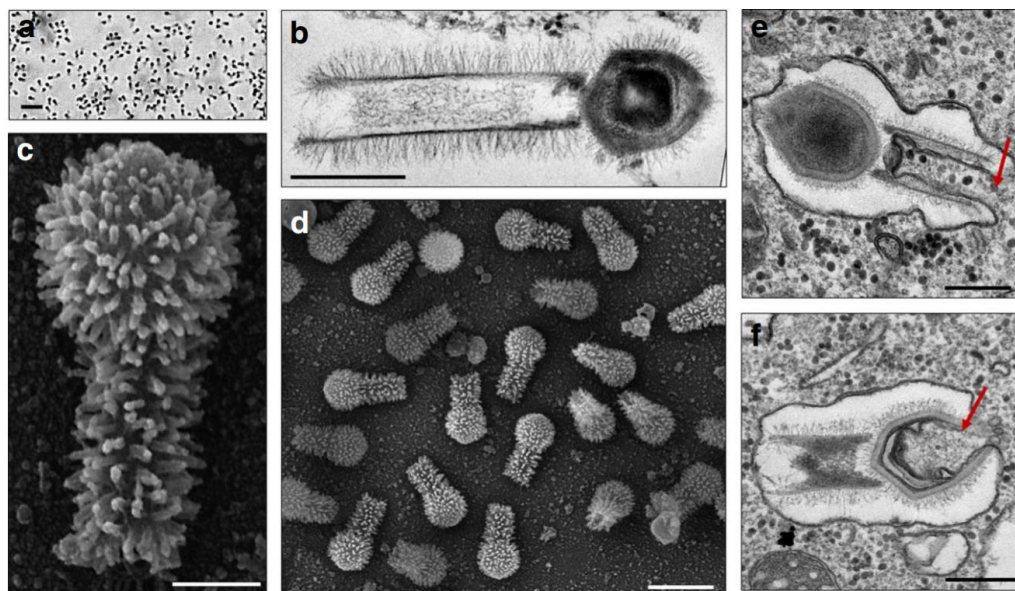


Figura 4. Tupanvírus soda lake, partículas e ciclo. (a) Microscopia óptica de partículas virais de tupanvírus após coloração por método hemacolor (1000 ×). Escala:

2 μm . (b) Super partícula ($>1000\text{ nm}$) observada por microscopia eletrônica de transmissão (TEM). Escala: 500 nm. (c, d) Microscopia eletrônica de varredura (MEV) de partículas de tupanvírus. Escala: 250 nm e 1 μm , respectivamente. (e, f) Fases iniciais da infecção em *A. castellanii*. Escala: 350 nm e 450 nm, respectivamente. Fonte: (ABRAHÃO et al., 2018) .

Após a descoberta dos vírus gigantes, que por si só já apresentam grande diversidade, a virosfera ficou ainda maior: foram descritos os primeiros virófagos – algo inimaginável algumas décadas atrás. Os virófagos são vírus satélites complexos associados a vírus gigantes e, portanto, infectam concomitantemente os hospedeiros dos vírus gigantes. A descoberta destes novos agentes virais contribuiu para que o conceito da identidade de um vírus passasse a ser novamente questionado (LA SCOLA et al., 2008).

1.2 O estudo de sequências

Desde o surgimento das primeiras tecnologias de sequenciamento genético, como o sequenciamento de Sanger, e principalmente conforme tais tecnologias evoluíram para uma maior capacidade de volume de dados gerados, diferentes aspectos a respeito de uma sequência de origem biológica podem ser alvos de estudo (PAREEK; SMOCZYNSKI; TRETYN, 2011; PARK; KIM, 2016).

Algumas das métricas primordiais de estudo de sequências tem como funcionalidade avaliar a qualidade de um sequenciamento em questão, como por exemplo métricas de quantificação de completude de sequência – que levam em conta o conteúdo sequenciado, como é o caso da ferramenta *BUSCO* – a fragmentação das sequências utilizando parâmetros estatísticos como o N50 – definição de N50 – dentre outros fatores de qualidade de sequenciamento, como a cobertura do mesmo (JAUHAL; NEWCOMB, 2021; MANNI et al., 2021).

Além da avaliação qualitativa do sequenciamento, existem ainda diversas outras abordagens para estudar sequências biológicas. Algumas das características que podem ser avaliadas e, a partir das quais já se é possível extrair importantes informações sobre o organismo de estudo, são: uso de códons, tendência de acúmulo de mutações, características específicas dos genes ou regiões codificadoras (do inglês “*coding DNA sequence*” ou CDS), característica das regiões entre genes, presença de elementos transponíveis de genoma, etc. As sequências biológicas, sendo elas DNA, RNA ou mesmo sequências proteicas, trazem em si, portanto, uma enorme riqueza de informação.

1.3 Estudos de composição de sequências biológicas e o conteúdo G+C

Dentre as inúmeras possibilidades de estudos em sequências, uma abordagem muito utilizada é dada pela observação e caracterização da composição das sequências, considerando as proporções dos ácidos nucleicos guanina (G), citosina (C), adenina (A) e timina (T) – no caso de sequências de DNA – ou uracila (U) – no caso de sequências de RNA. Estes trabalhos avaliam diferentes aspectos da composição de sequências, sendo um dos mais utilizados o percentual G+C, ou conteúdo GC%, que observa a proporção de guaninas e citosinas em uma determinada sequência. O cálculo utilizado para a obtenção desta informação é simples e se dá por $(G + C / A + T + G + C) \times 100\%$.

O conteúdo GC% é considerado uma das métricas mais simples de se analisar em uma sequência, mas que pode levar a informações biológicas de complexidade maior. Alguns estudos, por exemplo, utilizam-se do mapeamento do genoma de um organismo, correlacionando regiões de elevado GC% com a presença de *housekeeping genes*, ou

genes constitutivos para manutenção do funcionamento celular basal (SÉMON; MOUCHIROUD; DURET, 2005).

Outro contexto onde o estudo do perfil de conteúdo GC% em diferentes organismos traz importantes informações para a biologia destes é o caso da similaridade na composição de genomas de vírus e seus hospedeiros, como observado para vírus da família *Flaviviridae*, cujos hospedeiros são organismos vertebrados e invertebrados sob diferentes pressões evolutivas. Neste caso, o perfil de composição dos genomas destes vírus e de seus hospedeiros evidencia processos evolutivos linhagem-específicos compartilhados entre as espécies (LOBO et al., 2009).

Ainda, o perfil de composição de sequências e conteúdo GC% pode auxiliar na caracterização de microrganismos, auxiliando na identificação de isolados de um determinado grupo taxonômico, no qual este perfil é conservado, como é o caso de alguns poxvírus - isto porque a composição de sequências biológicas pode compor assinaturas genômicas linhagem-específicas. Estratégias como esta podem ser utilizadas, inclusive, na melhoria dos diagnósticos de infecções clínicas (LI et al., 2010). Tratando-se ainda de trabalhos de análise de sequências de microrganismos, o perfil de conteúdo GC% pode trazer informações sobre patogênese, estrutura e organização de genomas, perfil de mutação esperado e plasticidade genômica em relação a variações ambientais, dentre outros aspectos complexos que estão associados ao perfil de composição de sequências biológicas (KHRUSTALEV; BARKOVSKY, 2009; OUYANG et al., 2012; ROBINSON et al., 2013; WU et al., 2012).

1. JUSTIFICATIVA

Considerando o advento dos avanços tecnológicos das últimas décadas e, por consequência, a democratização e disponibilização de acesso a uma quantidade imensurável de dados públicos, dentre eles dados genômicos, é de se considerar estabelecer novas e mais diversas formas de se explorar este mar de conteúdo. Entendemos que é importante que as ferramentas e extração de informações biológicas não devem concentrar-se nas mãos de poucos, tampouco ficar restrita a metodologias complexas que dependem de uma elevada capacidade de processamento. Em vez disso, é possível que o material publicamente disponível nos grandes bancos de dados biológicos mundiais seja explorado, também, de maneira acessível a um maior contingente de pesquisadores.

Neste trabalho, especificamente, consideramos o grande grupo taxonômico da virologia, *Nucleocytoviricota* – um dos mais diversos e estudados grupos virais – um alvo de grande interesse de estudo. Possuindo integrantes de variação notável em tamanho de genoma, quantidade de regiões codificantes, nichos biológicos, organismos hospedeiros, dentre outras características muito diversas, os NCLDV's integram um grupo particularmente interessante para estudos comparativos.

Ainda, com a constante descoberta de novos vírus pertencentes ao grupo taxonômico, faz-se importante explorar todos os recursos disponíveis, para assim entender cada vez melhor a natureza da biologia dos NCLDV's. Neste estudo, aplicamos uma das métricas mais utilizadas em análises de composição de sequências biológicas, o conteúdo GC%, em uma abordagem direcionada a determinar o perfil de GC% das regiões codificadoras de NCLDV's. Esta abordagem tem o potencial de identificar padrões evolutivos em *Nucleocytoviricota* e dentro de famílias pertencentes ao filo. Além disto, é possível determinar, de acordo com a variação do conteúdo GC%, regiões alvo de interesse para trabalhos futuros.

2. OBJETIVOS

3.1 Objetivo geral

Descrever e analisar o perfil de conteúdo GC% ao longo do genoma de vírus grandes e gigantes pertencentes ao filo *Nucleocytoviricota*, levando em consideração cada região codificadora anotada e publicamente disponível de cada genoma selecionado.

3.2 Objetivos específicos

Os objetivos específicos deste trabalho foram:

1. Obter, por meio de ferramentas de bioinformática disponíveis publicamente, o valor percentual de GC referente a cada CDS anotada, em cada genoma;
2. Estabelecer uma estratégia semi-automatizada para a extração e análise de dados de GC%, e aplicá-la no presente estudo;
3. Descrever representações gráficas simples para visualização da oscilação nos valores de GC% atribuídos a cada CDS ao longo de um genoma;
4. Identificar padrões evolutivos entre os diferentes grupos do filo *Nucleocytoviricota*.

3. FLUXOGRAMA

Como representação simplificada da *pipeline* desenvolvida e executada na realização deste trabalho, encontra-se um fluxograma linear abaixo (**Figura 5**).

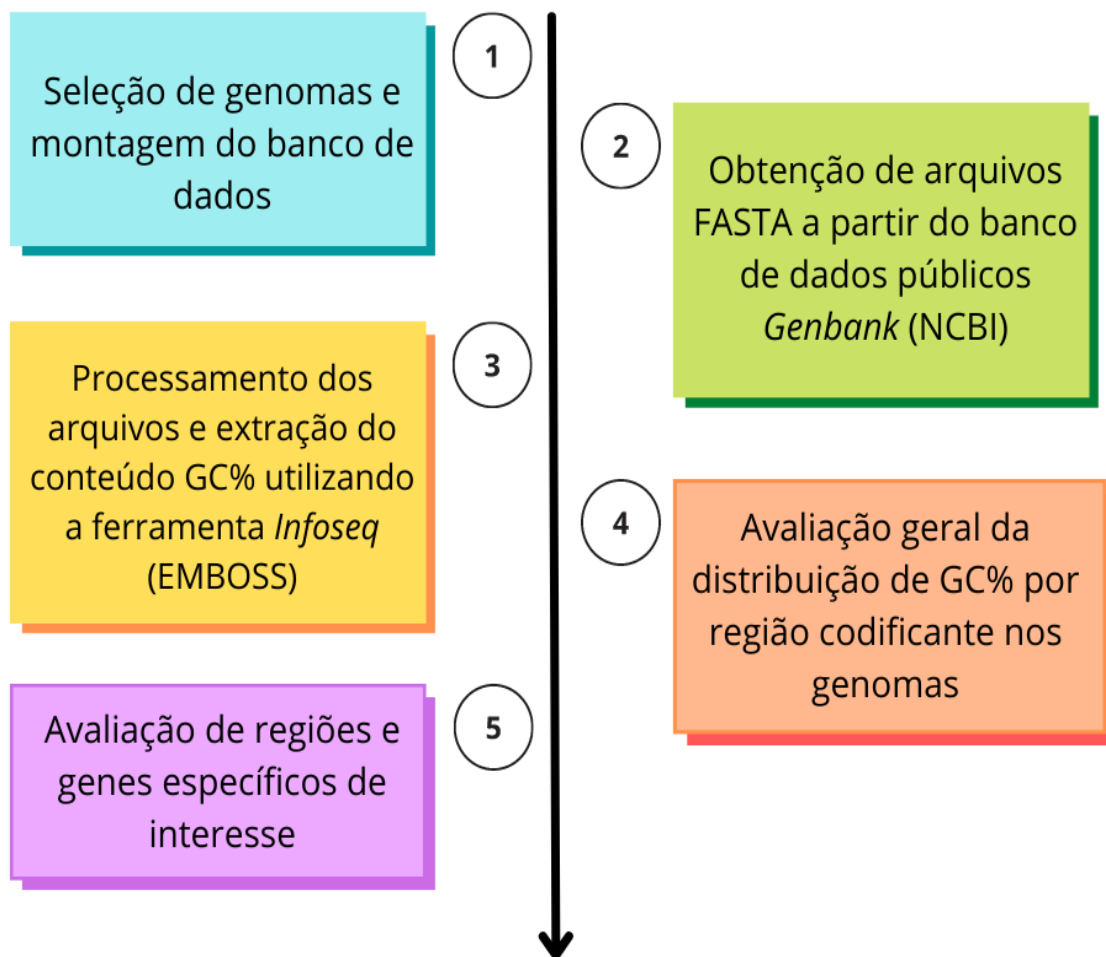


Figura 5. Fluxograma de esquema do trabalho realizado. (1) Seleção de genomas e montagem do banco de dados; (2) Obtenção de arquivos FASTA a partir do banco de dados públicos *Genbank* (NCBI); (3) Processamento dos arquivos e extração do conteúdo GC% utilizando a ferramenta *Infoseq* (EMBOSS); (4) Avaliação geral da distribuição de GC% por região codificante nos genomas e (5) Avaliação de regiões e genes específicos de interesse.

4. MATERIAL E MÉTODO

5.1 Seleção de genomas e aquisição de sequências

Como critérios para seleção dos genomas para o banco de dados, inicialmente foram escolhidos os vírus protótipos de seus respectivos subgrupos taxonômicos, em função da maior representatividade de estudos usualmente atrelados às espécies protótipos. No entanto, posteriormente, foram adicionados outros vírus que, independentemente desta classificação, possuíam relevância em seus respectivos grupos e/ou poderiam agregar variedade ao *pool* de genomas trabalhados, de acordo com nosso entendimento.

Para a realização deste trabalho, foram utilizados somente vírus cujos genomas estavam disponíveis completos e anotados no banco de dados genômicos públicos do *National Center for Biotechnology Information* (NCBI), o *GenBank* (BENSON *et al.*, 2000).

5.2 Processamento de genomas e determinação de conteúdo GC%

A partir da constituição de um banco de dados robusto contendo todos os integrantes do filo *Nucleocyotoviricota* a integrar este trabalho, prosseguimos para a aquisição das sequências de interesse. Assim, o genoma anotado de cada vírus em formato FASTA foi convertido em arquivos texto, constituído por uma lista de cada CDS anotada. Estes arquivos foram, então, submetidos à extração da informação de interesse: o conteúdo GC% referente a cada região codificadora anotada, em cada genoma.

Para a obtenção desta informação, utilizamos a ferramenta de processamento e análise de sequências biológicas, publicamente disponível e desenvolvida pelo *EMBL's European Bioinformatics Institute*, *EMBOSS* (disponível em <https://www.ebi.ac.uk/Tools/emboss/>). A partir do programa de análises bioinformáticas *EMBOSS*, o qual apresenta uma variedade de ferramentas de fácil acesso e utilização - mesmo para aqueles com menos experiência neste tipo de trabalho - utilizamos a ferramenta *InfoSeq*.

Com esta ferramenta, foi possível obter informações objetivas acerca de cada sequência. Neste caso, especificamente, os *outputs* gerados pelo programa foram arquivos texto contendo uma lista informativa do nome da região codificadora, seguida do tamanho relativo a esta sequência e, finalmente, o valor do conteúdo GC% referente à mesma. Desta forma, obtivemos o GC% relativo a cada CDS anotada individualmente, para todos os genomas trabalhados.

5.3 Avaliação de GC% gene a gene

Com a intenção de melhor compreender a dimensão e o perfil da variação do conteúdo GC%, gene a gene, ao longo dos genomas virais estudados, foram plotados gráficos de dispersão – que foram adequadamente alocados em painéis referentes a cada grupo taxonômico – e, também, foram elaborados manualmente gráficos similares a *heatmaps* utilizando a ferramenta Excel da Microsoft.

Tais representações gráficas foram escolhidas para que visualmente fosse possível identificar padrões e/ou regiões e CDS específicas, cuja porcentagem de GC fosse suficientemente destoante do restante do genoma, indicando prováveis alvos de interesse para estudos aprofundados.

5.3.1 Construção de painéis de dispersão

As representações gráficas em painéis de dispersão foram feitas utilizando a ferramenta de gráficos do Microsoft Excel e, a partir dos gráficos sugeridos, um modelo foi escolhido e normalizado para este trabalho. Na construção dos gráficos de cada genoma, os valores percentuais de GC referente a cada CDS foram indicados no eixo Y (ordenadas), enquanto cada CDS possui uma posição relativa representada no eixo X (abscissas). Além disso, a linha de tendência média dos valores de cada genoma foi exposta nos respectivos gráficos, em vermelho.

Para a elaboração dos painéis, levando em consideração que um mesmo grupo taxonômico abrangia genomas de tamanhos muito variados e, conseqüentemente, quantidades de CDS a serem representadas graficamente muito variadas, foi estipulada a normalização do tamanho dos gráficos em função do maior genoma do grupo. Desta forma, a disposição dos genomas nos painéis respeita e ilustra, em algum nível, as diferenças de tamanho e quantidade de genes destes genomas virais.

5.3.2. Construção dos mapas de variação de GC%

Os mapas de variação de GC%, idealizados a partir de *heatmaps* usualmente utilizados em trabalhos de genômica, foram desenvolvidos na plataforma do Microsoft Excel, assim como os painéis de dispersão. Para realização de tal, foram inseridos em planilhas separadas os genomas de vírus pertencentes a cada grupo representado (*Phycodnaviridae*, *Mimiviridae*, *Marseilleviridae*, *Iridoviridae*, *Asfarviridae*, *Poxviridae*, “*Pithoviridae*” e outros *Nucleocytoviricota*) da seguinte forma: cada genoma ocupava uma coluna, na qual foram dispostos os valores de GC% calculado para cada CDS, na ordem em que foram anotadas e depositadas no *GenBank*.

Desta forma, em uma coluna (um genoma viral), cada linha representava um valor de GC% (referente a uma CDS). De posse destas planilhas, foram inseridas formatações condicionais de 3 formas, como descritas a seguir:

- A. Cada coluna foi colorizada de modo que células com valor médio de GC% do genoma representado ganham a cor amarela, células com valor próximo do valor máximo de GC% do genoma ganham cor vermelha e células com valor próximo do valor mínimo de GC% do genoma ganham cor azul;
- B. Todas as colunas foram coloridas numa mesma escala, diferente de A, desta vez utilizando o valor médio de GC% das CDS do grupo como referência de média em cor amarela, valor máximo e mínimo de GC% do grupo com cores vermelho e azul, respectivamente;
- C. Similar ao que foi feito em B, todas as colunas foram coloridas na mesma escala, mas foram utilizados valores médios, ainda coloridos em amarelo, levando em conta os genomas totais e não somente os valores de GC% das CDS.

Para ocultar os valores numéricos das células coloridas, afim de obter imagens com melhor resolução, adicionou-se a configuração de exibição de números no Excel de

forma personalizada, correspondente a “;””. Esta configuração permite que o Excel exiba células visualmente “vazias”, enquanto esconde e mantém os respectivos valores numéricos inseridos.

5. RESULTADOS

6.1 Banco de dados e genomas utilizados

Durante a montagem do banco de dados a ser utilizado como material de estudo deste trabalho, foram selecionados um total de 61 genomas virais pertencentes ao grupo taxonômico de *Nucleocytoviricota*. Destes, foram selecionados, através de uma extensa curadoria manual, as seguintes quantidades de genomas e suas famílias taxonômicas correspondentes: quatro genomas pertencentes à *Phycodnaviridae*, 13 genomas pertencentes à *Mimiviridae*, um genoma de *Ascoviridae*, cinco de *Iridoviridae*, cinco de *Marseilleviridae*, um de *Asfarviridae*, 22 de *Poxviridae*, e 10 genomas de Unclassified DNA vírus, incluindo exemplares de “Pithoviridae”, “Medusaviridae”, *Yaraviridae*, “Pandoraviridae” e “extended Asfarviridae” (**tabela 1**).

Para os devidos fins de consulta ao número de cadastro de patrimônio genético dos agentes biológicos estudados neste trabalho, estes estão cadastrados sob os seguintes acessos do SisGen (Sistema Nacional de Gestão do Patrimônio Genético e do Conhecimento Tradicional Associado): AA3B21E, ABF23CC, A702EB8, ABF23CC, A25764F, ABF23CC, AC31840, ABF23CC, A3DAB3F, ABF23CC, AC3045D, ABF23CC, A98DD8C, AC3045D, ABF23CC, ABF23CC e AB9AB93.

Tabela 1. Total de 61 genomas virais pertencentes ao filo *Nucleocyotoviricota* selecionados e utilizados neste trabalho. Estão representados junto aos seus respectivos números de acesso ao banco de dados *GenBank*, tamanho do genoma em Kb e valores médios de GC% de cada genoma total e de GC% de cada genoma considerando somente as regiões codificadoras.

Classificação taxonômica	Vírus	GenBank Accession	Tamanho do genoma (Kb)	GC% médio total	GC% médio das CDS
<i>Phycodnaviridae</i>	Paramecium bursaria Chlorella virus 1	NC_000852.5	330,61	40,0	40,6
<i>Phycodnaviridae</i>	Emiliana huxleyi virus 86	NC_007346.1	407,34	40,2	40,6
<i>Phycodnaviridae</i>	Ectocarpus siliculosus virus 1	NC_002687.1	335,69	51,7	52,2
<i>Phycodnaviridae</i>	Heterosigma akashiwo virus 1	NC_038553.1	274,79	30,4	30,8
<i>Mimiviridae</i>	Cafeteria roenbergensis virus	NC_014637.1	617,45	23,3	23,4
<i>Mimiviridae</i>	Acanthamoeba polyphaga mimivirus	NC_014649.1	1181,55	28,0	28,8
<i>Mimiviridae</i>	Samba virus	KF959826.2	1181,55	28,0	28,7
<i>Mimiviridae</i>	Megavirus chiliensis	NC_016072.1	1246,13	25,3	26,3
<i>Mimiviridae</i>	Moumouvirus australiensis	MG807320.1	1098	25,1	26,0
<i>Mimiviridae</i>	Tupanvirus deep ocean	MF405918.2	1439,51	29,4	30,5
<i>Mimiviridae</i>	Tupanvirus soda lake	KY523104.2	1516,27	29,1	30,2
<i>Mimiviridae</i>	Aureococcus anophagefferens virus	NC_024697.1	370,92	28,7	29,3
<i>Mimiviridae</i>	Bodo Saltans Virus	MF782455.1	1385,87	25,3	25,7
<i>Mimiviridae</i>	Phaeocystis globosa virus group I	NC_021312.1	459,98	32,0	33,4
<i>Mimiviridae</i>	Chrysochromulina ericina virus	NC_028094.1	473,56	25,4	26,0
<i>Mimiviridae</i>	Tetraselmis viridis virus	KY322437.1	668,03	41,2	40,6
<i>Mimiviridae</i>	Cotonvirus japonicus	AP024483.1	1476,527	25,3	26,6
<i>Ascoviridae</i>	Spodoptera frugiperda ascovirus 1a	NC_008361.1	156,92	49,2	50,3
<i>Iridoviridae</i>	Lymphocystis disease virus 1	NC_001824.1	102,65	29,1	29,2
<i>Iridoviridae</i>	Frog virus 3	NC_005946.1	105,903	55,1	57,1
<i>Iridoviridae</i>	Invertebrate iridescent virus 3	NC_008187	191,1	47,9	50,4
<i>Iridoviridae</i>	Decapod iridescent virus 1	MF599468.1	165,81	34,6	29,8
<i>Iridoviridae</i>	Invertebrate iridescent virus 6	NC_003038.1	212,48	28,6	35,0
<i>Marseilleviridae</i>	Marseillevirus marseillevirus	NC_013756.1	368,45	44,7	45,0
<i>Marseilleviridae</i>	Brasillian marseillevirus	KT752522	362,28	43,3	43,9
<i>Marseilleviridae</i>	Golden marseillevirus	NC_031465.1	360,61	43,1	43,7
<i>Marseilleviridae</i>	Lausannevirus	NC_015326.1	346,75	42,9	43,0
<i>Marseilleviridae</i>	Tunisvirus	NC_038511.1	380,01	43,0	43,6
<i>Asfarviridae</i>	African swine fever virus strain BA71V	NC_001659.2	189,37	38,6	38,7
Extended Asfarviridae	Pacmanvirus S19	MZ440852.1	418,588	33,2	34,3
Extended Asfarviridae	Faustovirus e12	KJ614390	465,956	37,7	36,9
Extended Asfarviridae	Kaumoebavirus	MT334784.1	362,586	43,1	43,4
<i>Poxviridae</i>	Fowlpox virus	NC_002188.1	291,058	31,2	31,3
<i>Poxviridae</i>	Sheeppox virus 17077-99	NC_004002.1	149,956	25,0	25,3
<i>Poxviridae</i>	Yokapox virus	NC_015960.1	175,7	25,6	26,2
<i>Poxviridae</i>	Mule deerpox virus	AY689437.1	170,56	27,0	27,6
<i>Poxviridae</i>	Nile crocodilepox virus	NC_008030.1	190,05	61,9	62,4
<i>Poxviridae</i>	Myxoma virus	NC_001132.2	162,455	43,5	43,5
<i>Poxviridae</i>	Eastern kangaroopox virus	MF467281.1	170,141	54,0	54,3
<i>Poxviridae</i>	Molluscum contagiosum virus	MH646551.1	192,156	64,3	63,9
<i>Poxviridae</i>	Sea otterpox virus	NC_037656.1	127,88	31,3	31,6
<i>Poxviridae</i>	Vaccinia virus	NC_006998.1	182,546	33,4	34,6
<i>Poxviridae</i>	Cotia virus	KM595078.1	185,14	23,6	24,4
<i>Poxviridae</i>	Orf virus	NC_005336.1	139,962	63,8	64,3
<i>Poxviridae</i>	Pteropox virus	NC_030656.1	133,49	33,8	34,0
<i>Poxviridae</i>	Salmon gillpox virus	NC_027707.1	241,56	37,5	37,1
<i>Poxviridae</i>	Squirrelpox virus	NC_022563.1	148,8	66,7	67,0
<i>Poxviridae</i>	Swinepox virus	NC_003389.1	146,454	27,4	27,8
<i>Poxviridae</i>	Eptesipox virus	NC_035460.1	176,69	23,6	23,9
<i>Poxviridae</i>	Yaba monkey tumor virus	NC_005179.1	134,72	29,8	30,1
<i>Poxviridae</i>	Anomala cuprea entomopoxvirus	NC_023426.1	245,72	20,0	20,5
<i>Poxviridae</i>	Amsacta moorei entomopoxvirus	NC_002520.1	232,39	17,8	18,2
<i>Poxviridae</i>	Melanoplus sanguinipes entomopoxvirus	NC_001993.1	236,12	18,3	19,0
<i>Poxviridae</i>	Diachasmimorpha longicaudata entomopoxvirus	KR095315.1	252,94	30,1	31,0
<i>Pithoviridae</i>	Pithovirus sibericum	NC_023423.1	610,03	35,8	40,2
<i>Pithoviridae</i>	Cedratvirus A11	NC_032108.1	589,07	42,7	43,0
Pithoviridae-like	Orpheovirus	NC_036594.1	1,473,57	25,0	28,1
Yaravirus	Yaravirus brasiliensis	MT293574.1	44,924	58,0	58,0
Pandoravirus	Pandoravirus quercus	NC_037667.1	2077,29	60,7	64,4
Unclassified virus	Medusavirus	AP018495.1	381,277	61,7	61,5
Unclassified virus	Mollivirus sibericum	NC_027867.1	651,52	60,1	60,2

6.2 Variação do perfil de conteúdo GC%

Em *Phycodnaviridae* (**figura 6**), estão representados quatro genomas, sendo estes *Paramecium bursaria Chlorella vírus 1*, *Emiliana huxleyi vírus 86*, *Heterosigma akashiwo vírus 1* e *Ectocarpus siliculosus vírus 1*. Destes, o genoma com maior quantidade de CDS anotadas é o de *Paramecium bursaria Chlorella vírus 1*, enquanto o genoma com menos CDS anotadas é o de *Ectocarpus siliculosus virus 1*.

As linhas de tendência médias de *Paramecium bursaria Chlorella vírus 1* e *Emiliana huxleyi vírus 86* estão representadas em torno do valor 40 de percentual GC, enquanto os demais genomas possuem tendências médias discrepantes, sendo de aproximadamente 30% em *Heterosigma akashiwo vírus 1* e aproximadamente 50 em *Ectocarpus siliculosus virus 1*.

Em relação à dispersão dos valores GC% ao longo dos genomas representados, não observamos nenhum padrão específico. No entanto, o genoma de *Emiliana huxleyi vírus 86* apresentou uma característica distinta: um agrupamento de genes – ou *cluster* – com valores significativamente mais elevados em relação à média do genoma, aproximado da posição relativa 300.



Figura 6. Representação da dispersão dos valores de conteúdo GC%, gene a gene, em genomas de Phycodnaviridae. Representação gráfica da dispersão dos valores de percentual GC (identificados no eixo Y) para cada CDS anotada disponível (identificadas pelos pontos posicionados na extensão do eixo X), para os genomas de *Paramecium bursaria* a *Chlorella* vírus 1, *Emiliana huxleyi* vírus 86, *Heterosigma akashiwo* vírus 1 e *Ectocarpus siliculosus* vírus 1. Linha de tendência dos valores médios representada em vermelho, sobre os valores pontuais de GC% de cada genoma.

Já em *Mimiviridae* (**figuras 7 e 8**), os genomas com maior e menor número de CDS anotadas são de tupanvirus soda lake e Aureococcus anophagefferens vírus, respectivamente. As linhas de tendência média de valores percentuais GC em *Mimiviridae* estão representadas, majoritariamente, em torno de valores baixos entre 20% e 30% com exceção do genoma de Tetraselmis viridis vírus, cuja linha de tendência de valores GC% encontra-se acima do marco 40%. Em relação ao perfil de dispersão de valores GC% neste grupo, também não foi observado nenhum padrão específico.

Os genomas representados possuem valores de conteúdo GC% por CDS dispostos próximos das respectivas médias, em sua maioria, contando com *outliers* – principalmente com valores significativamente maiores do que o valor médio correspondente. Os genomas de Phaeocystis globosa vírus, Chrysochromulina ericina vírus, megavirus chiliensis e cotonvirus japonicus são alguns dos que mais destacam-se pela identificação de *outliers*.

Ainda, foram observados genomas com conjuntos de CDS, ou clusters de regiões codificadoras, para os quais o valor GC% é maior do que as respectivas médias, sendo estes Chrysochromulina ericina vírus (em torno da posição relativa 450), Cafeteria roenbergensis vírus (em torno da posição relativa 100), Aureococcus anophagefferens vírus (em torno da posição relativa 10) e cotonvirus japonicus (dois picos observados em torno da posição relativa 480). Bodo saltans vírus, por sua vez, apresentou ambas as regiões terminais do genoma com valores GC% majoritariamente acima de sua média.

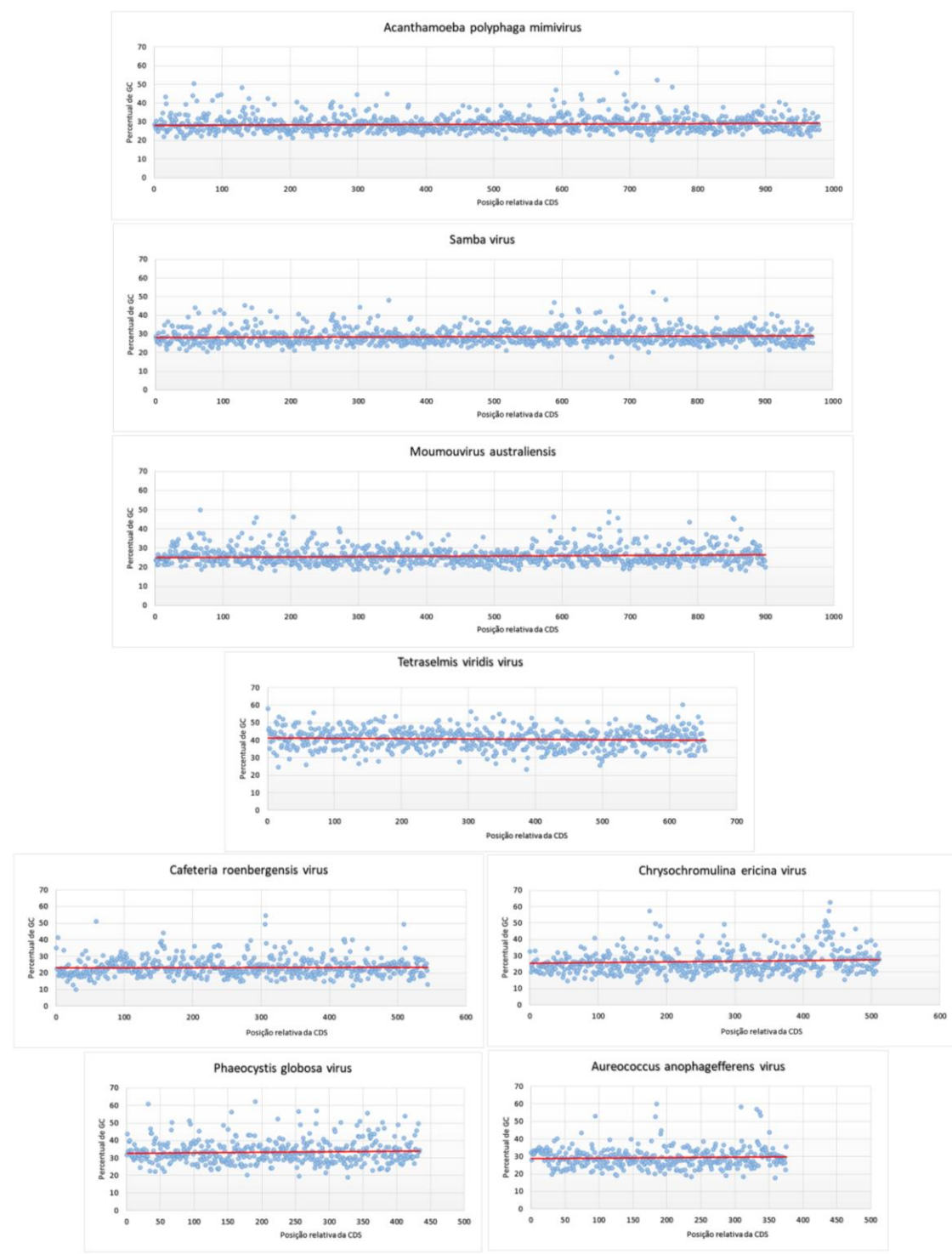


Figura 7. Representação da dispersão dos valores de conteúdo GC%, gene a gene, em genomas de Mimiviridae (painel A). Representação gráfica da dispersão dos valores de percentual GC (identificados no eixo Y) para cada CDS anotada disponível

(identificadas pelos pontos posicionados na extensão do eixo X), para os genomas de *Acanthamoeba polyphaga* mimivirus, Samba vírus, mousmouvirus australiensis, Tetraselmis viridis vírus, Cafeteria roenbergensis vírus, Chrysochromulina ericina vírus, Phaeocystis globosa vírus e Aureococcus anophagefferens vírus. Linha de tendência dos valores médios representada em vermelho, sobre os valores pontuais de GC% de cada genoma.

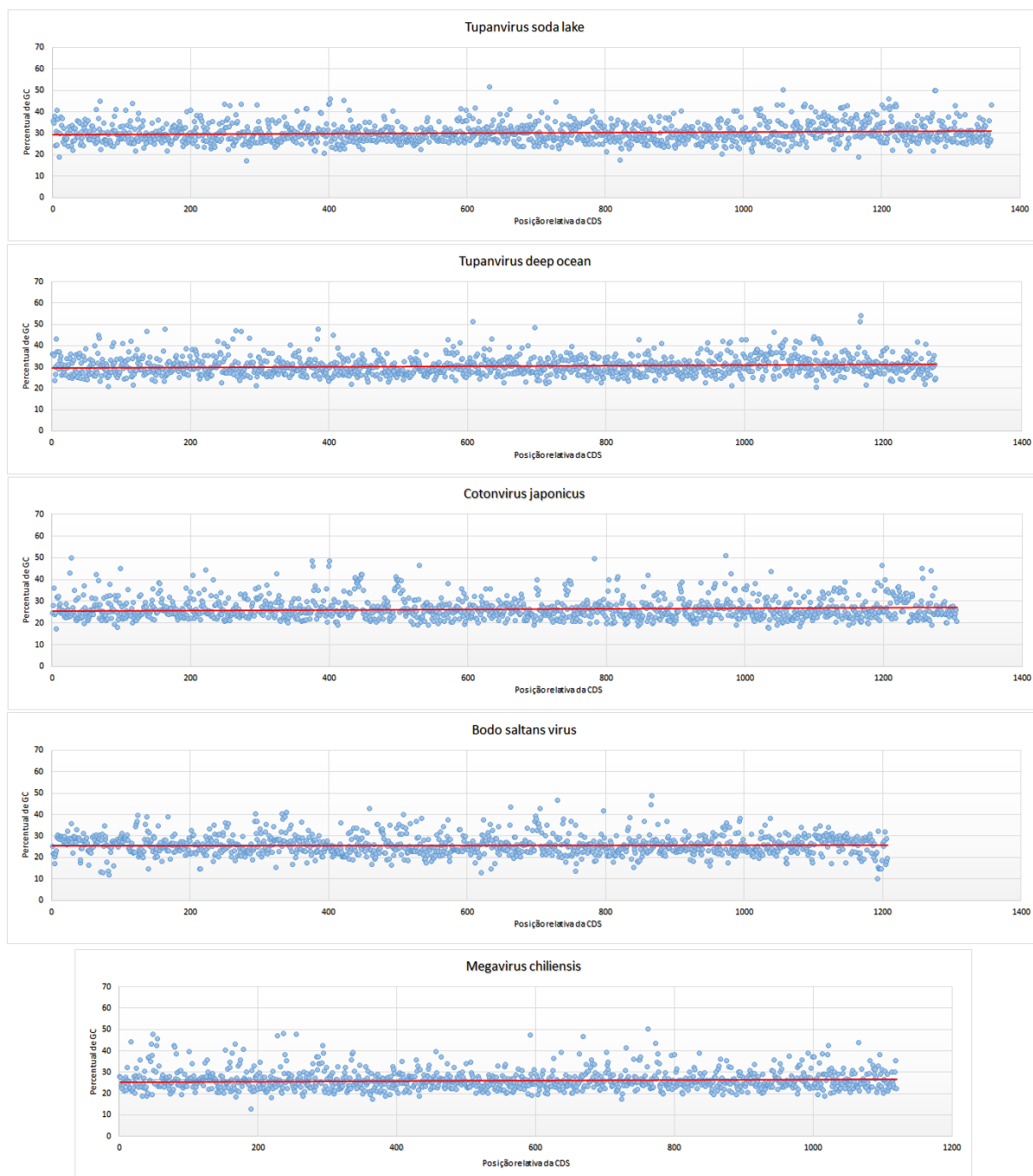


Figura 8. Representação da dispersão dos valores de conteúdo GC%, gene a gene, em genomas de Mimiviridae (painel B). Representação gráfica da dispersão dos valores de percentual GC (identificados no eixo Y) para cada CDS anotada disponível (identificadas pelos pontos posicionados na extensão do eixo X), para os genomas de *tupanvirus soda lake*, *tupanvirus deep ocean*, *cotonvirus japonicus*, *Bodo saltans vírus* e

megavirus chiliensis. Linha de tendência dos valores médios representada em vermelho, sobre os valores pontuais de GC% de cada genoma.

Em membros da família *Iridoviridae* (**figura 9**), os genomas com maior e menor número de CDS anotadas foram, respectivamente, Invertebrate iridescente vírus 6 com 468 proteínas preditas e Frog vírus 3 com 99 proteínas preditas.

As linhas de tendência média dos valores de GC% observados nesse grupo estão representadas em valores aproximados de 30%, para os vírus Invertebrate iridescente vírus 6, Lymphocystis Disease vírus 1 e Decapod iridescent vírus 1, ou valores aproximados de 50%, para os genomas de Frog vírus 3 e Invertebrate iridescente vírus 3.

O perfil de dispersão dos valores GC% em *Iridoviridae* não apresenta padrões específicos, além de uma breve tendência a valores GC% mais aproximados das relativas médias individuais e presença de poucos *outliers*.

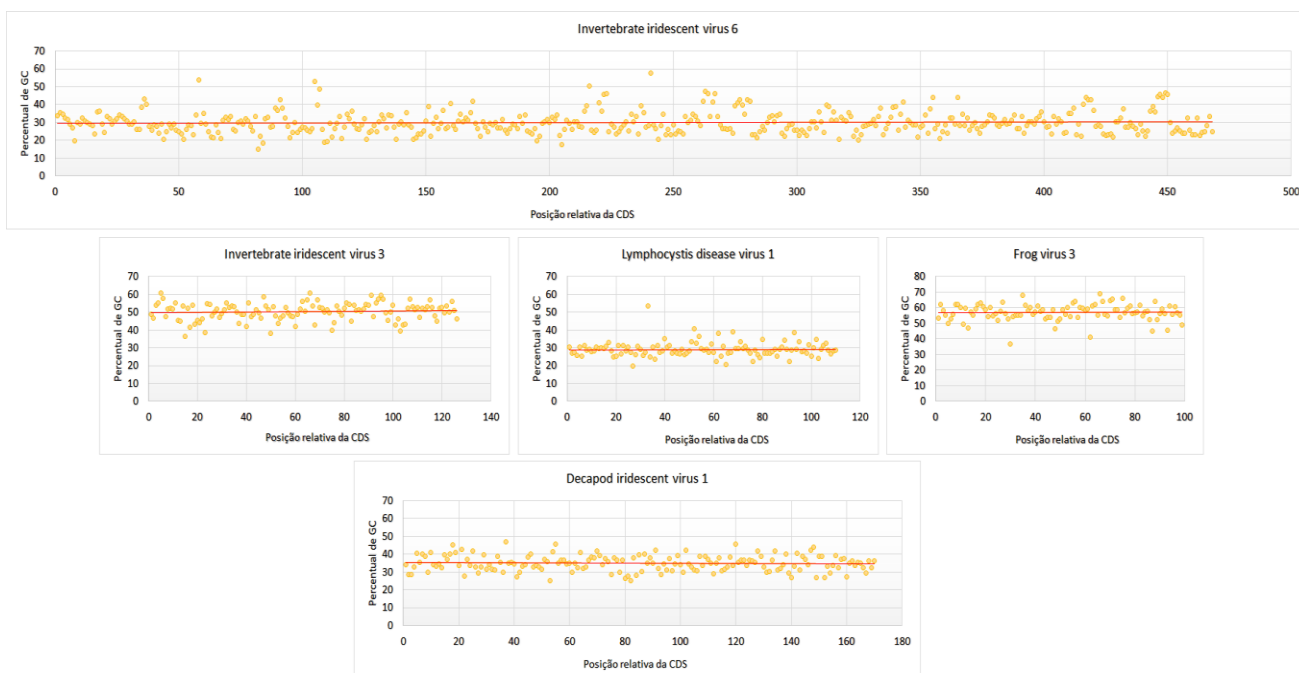


Figura 9. Representação da dispersão dos valores de conteúdo GC%, gene a gene, em genomas de Iridoviridae. Representação gráfica da dispersão dos valores de percentual GC (identificados no eixo Y) para cada CDS anotada disponível (identificadas pelos pontos posicionados na extensão do eixo X), para os genomas de Invertebrate iridescente vírus 6, Invertebrate iridescente vírus 3, Lymphocystis Disease vírus 1, Frog vírus 3 e Decapod iridescente vírus 1. Linha de tendência dos valores médios representada em vermelho, sobre os valores pontuais de GC% de cada genoma.

Nos vírus pertencentes à família *Marseilleviridae* (**figura 10**), contando com cinco genomas com mínima de 296 CDS anotadas – golden marseillevírus – e máxima de 484 CDS anotadas – tunisvirus – observamos uma consistência em relação às linhas de tendência de valor médio de GC%, concentrando-se em aproximadamente 45%.

Além desta característica conservada entre os cinco genomas estudados, observamos, também, uma tendência de valores GC% muito próximos às relativas médias individuais, com a presença de poucos *outliers*.

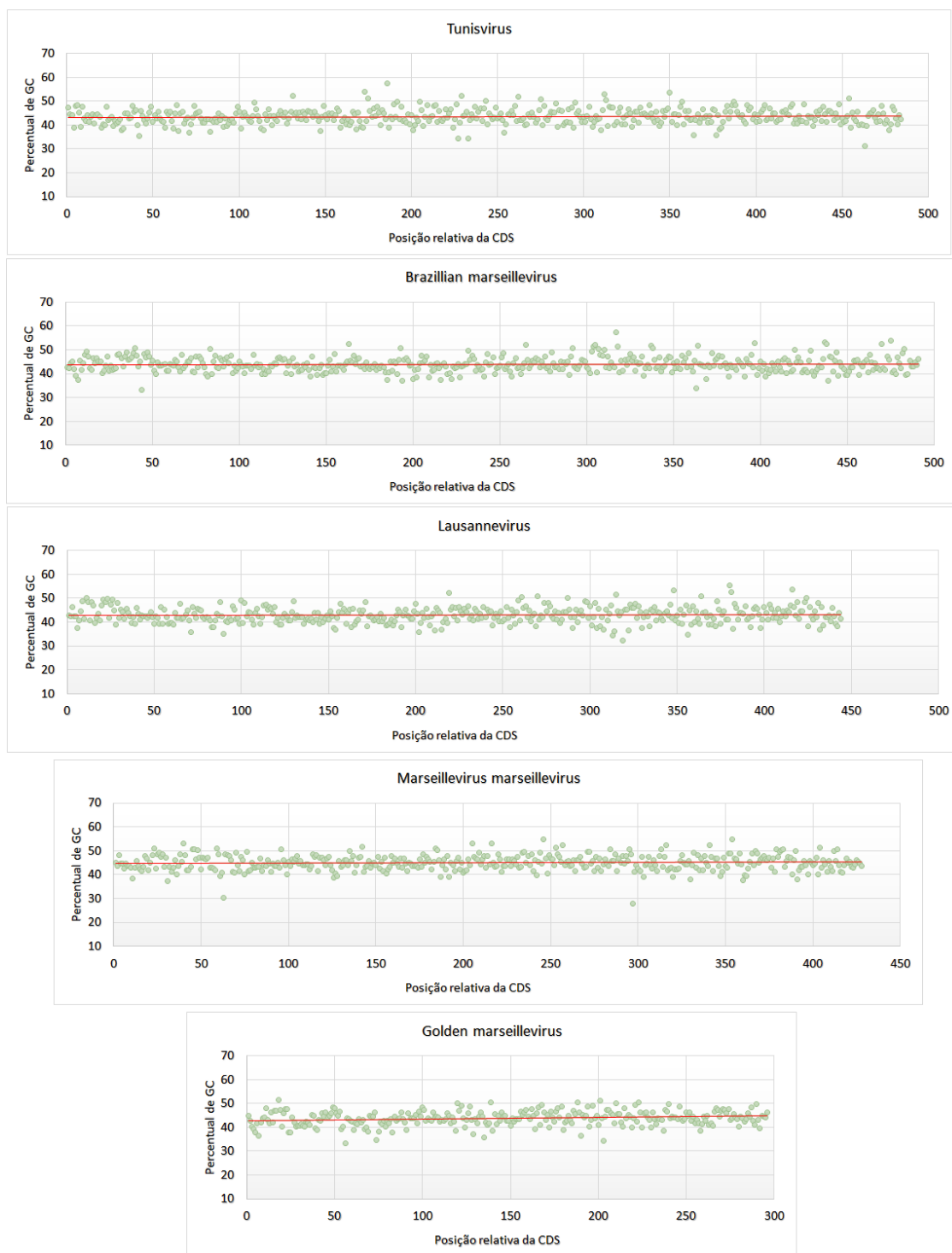


Figura 10. Representação da dispersão dos valores de conteúdo GC%, gene a gene, em genomas de *Marseilleviridae*. Representação gráfica da dispersão dos valores de percentual GC (identificados no eixo Y) para cada CDS anotada disponível

(identificadas pelos pontos posicionados na extensão do eixo X), para os genomas de *tunisvirus*, *brazilian marseillevirus*, *lausannevirus*, *marseillevirus* *marseillevirus* e *golden marseillevirus*. Linha de tendência dos valores médios representada em vermelho, sobre os valores pontuais de GC% de cada genoma.

Em vírus pertencentes à família *Asfarviridae* (**figura 11**), os genomas com maior e menor número de CDS anotadas são de *kaumoebavirus* e *african swine fever virus*, com 508 e 153 regiões codificadoras, respectivamente.

Neste grupo, embora as linhas de tendências médias de valor GC% estejam representadas em posições relativamente próximas, entre 30% e 45%, a dispersão dos valores GC% ao longo dos respectivos genomas é bastante variada. Foram observados genomas com GC% constante e próximo da média individual do genoma – como é o caso, principalmente, de *kaumoebavirus* – e genomas cujos valores GC% apresentavam grandes amplitudes, como em *pacmanvirus* S19 e *african swine fever virus*.

Além destas características, não foram observadas tendências compartilhadas no grupo taxonômico.

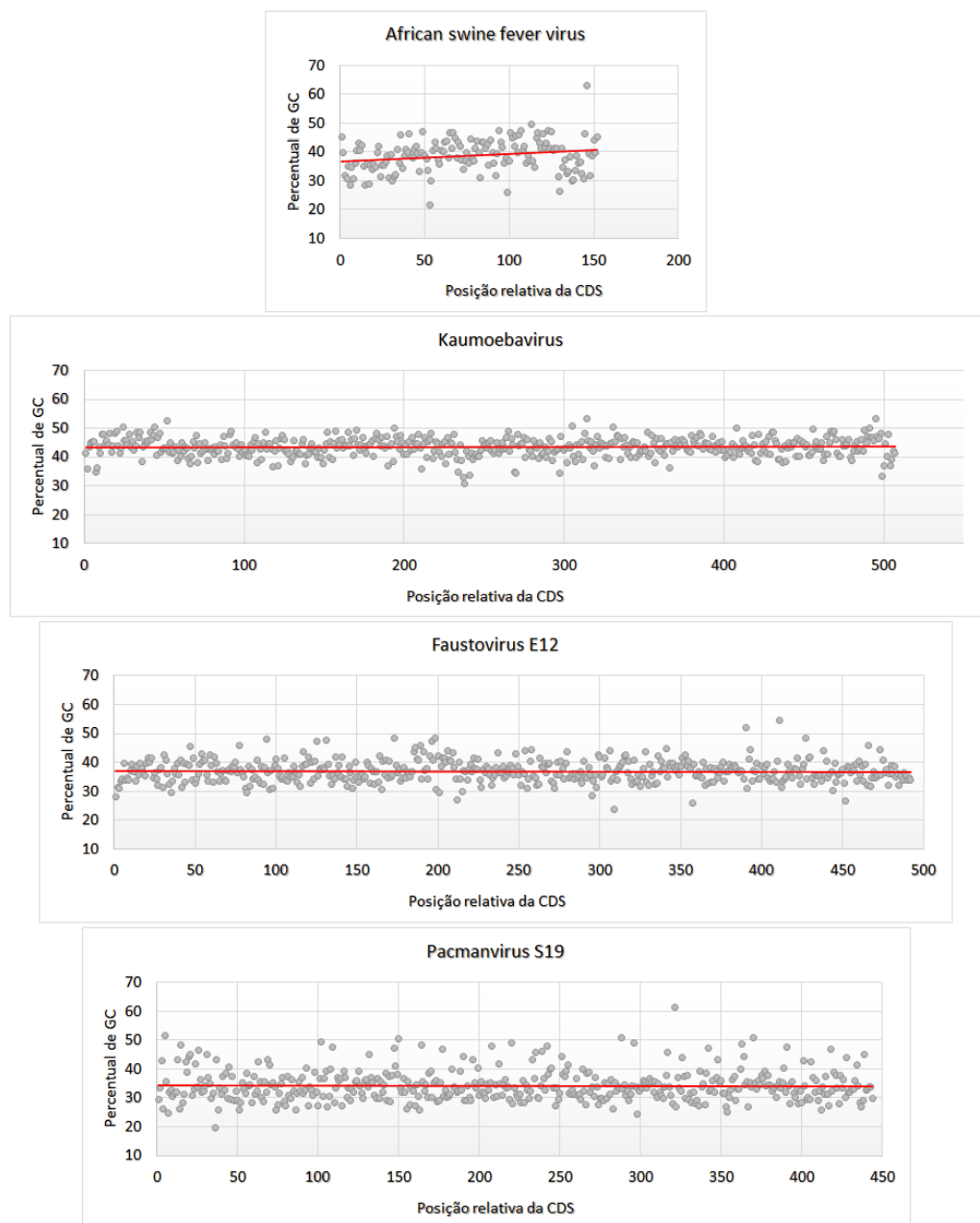


Figura 11. Representação da dispersão dos valores de conteúdo GC%, gene a gene, em genomas de *Asfarviridae*. Representação gráfica da dispersão dos valores de percentual GC (identificados no eixo Y) para cada CDS anotada disponível (identificadas pelos pontos posicionados na extensão do eixo X), para os genomas de african swine fever vírus, kaumobavirus, faustovirus E12 e pacmanvirus S19. Linha de tendência dos valores médios representada em vermelho, sobre os valores pontuais de

GC% de cada genoma. Neste painel, “extended asfarviridae” foi representada em conjunto a *Asfarviridae*.

Em *Poxviridae* (**figuras 12 e 13**), desconsiderando a subfamília *Entomopoxvirinae*, o maior e menor número de CDS observados foram de 261 e 130, em fowlpoxvirus e orf vírus, respectivamente. Para *Entomopoxvirinae*, observamos *Amsacta moorei* entomopoxvirus, com 293 proteínas preditas, e *Diachasmimorpha longicaudata* entomopoxvirus, com 193 proteínas preditas.

De modo geral, *Poxviridae* foi o grupo taxonômico dentro de *Nucleocyotoviricota* que apresentou maior amplitude em relação aos valores de GC% atribuídos a uma CDS específica, e também em relação às médias de GC% individuais dos genomas – valores a serem discutidos nas próximas seções.

Apesar da grande variabilidade de valores GC% em *Poxviridae*, o grupo apresentou majoritariamente genomas com baixa amplitude de valores GC%, que se concentram próximos aos respectivos valores médios – com a exceção de nile crocodilepox virus. Em *Entomopoxvirinae* (**figura 14**), no entanto, observamos genomas com variação de GC% discretamente maior, como no caso de *Amsacta moorei* entomopoxvirus.

Ainda, os genomas em *Poxviridae* apresentaram uma outra característica relevante em relação à dispersão de valores GC% associados a regiões codificadoras – valores GC% condicionados à CDS nas extremidades dos genomas se mantinham, de modo geral, semelhantes e inversamente representados ao decorrer do genoma (isto é, espelhados). Em alguns dos gráficos referentes a poxvírus, como é o caso de vaccinia vírus e myxoma vírus, esta característica fica em maior evidência.

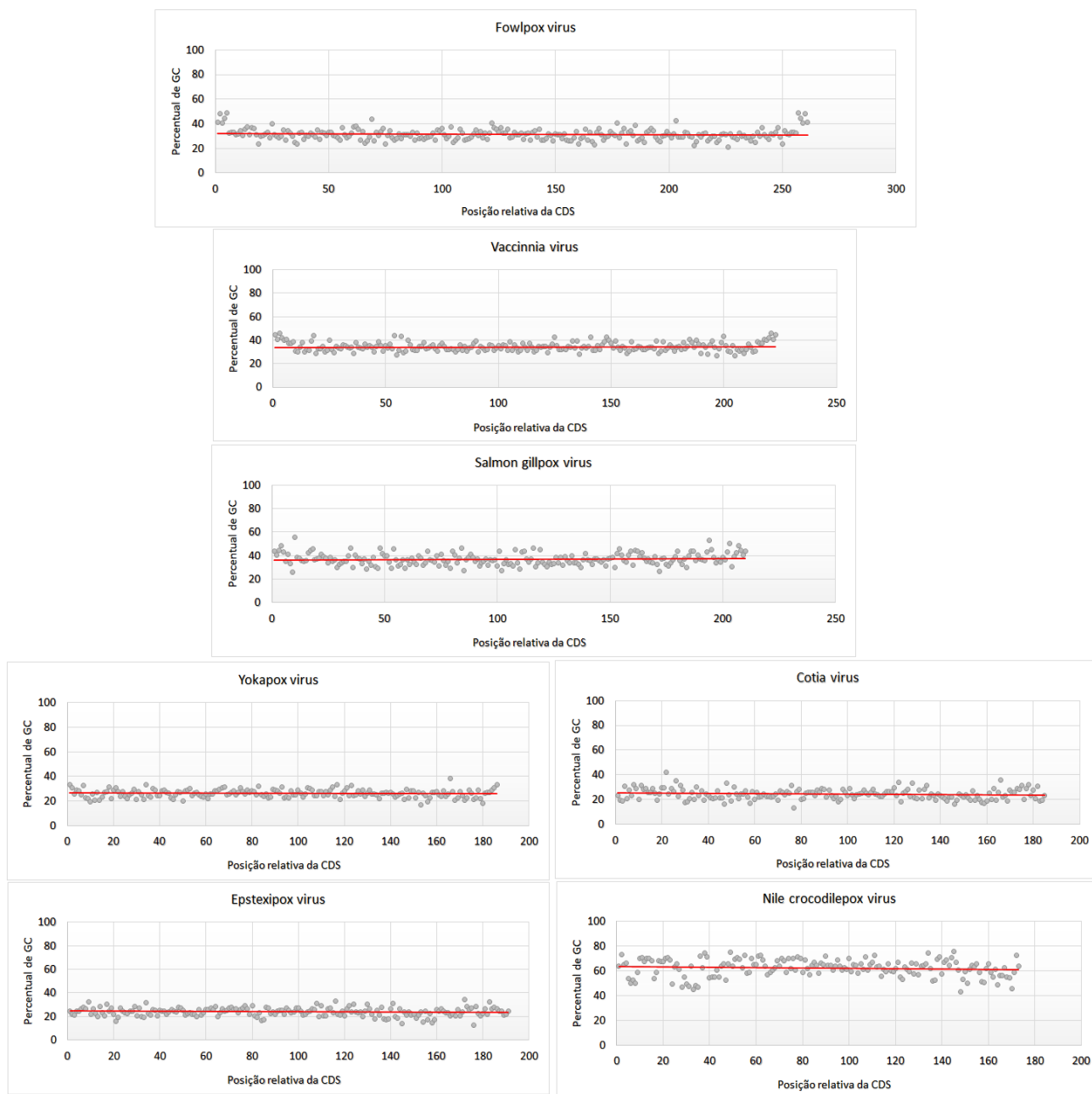


Figura 12. Representação da dispersão dos valores de conteúdo GC%, gene a gene, em genomas de Poxviridae (painel A). Representação gráfica da dispersão dos valores de percentual GC (identificados no eixo Y) para cada CDS anotada disponível (identificadas pelos pontos posicionados na extensão do eixo X), para os genomas de fowlpox vírus, vaccinia vírus, salmon gillpox vírus, yokapox vírus, cotia vírus, epstesipox

vírus e Nile crocodilepox vírus. Linha de tendência dos valores médios representada em vermelho, sobre os valores pontuais de GC% de cada genoma.

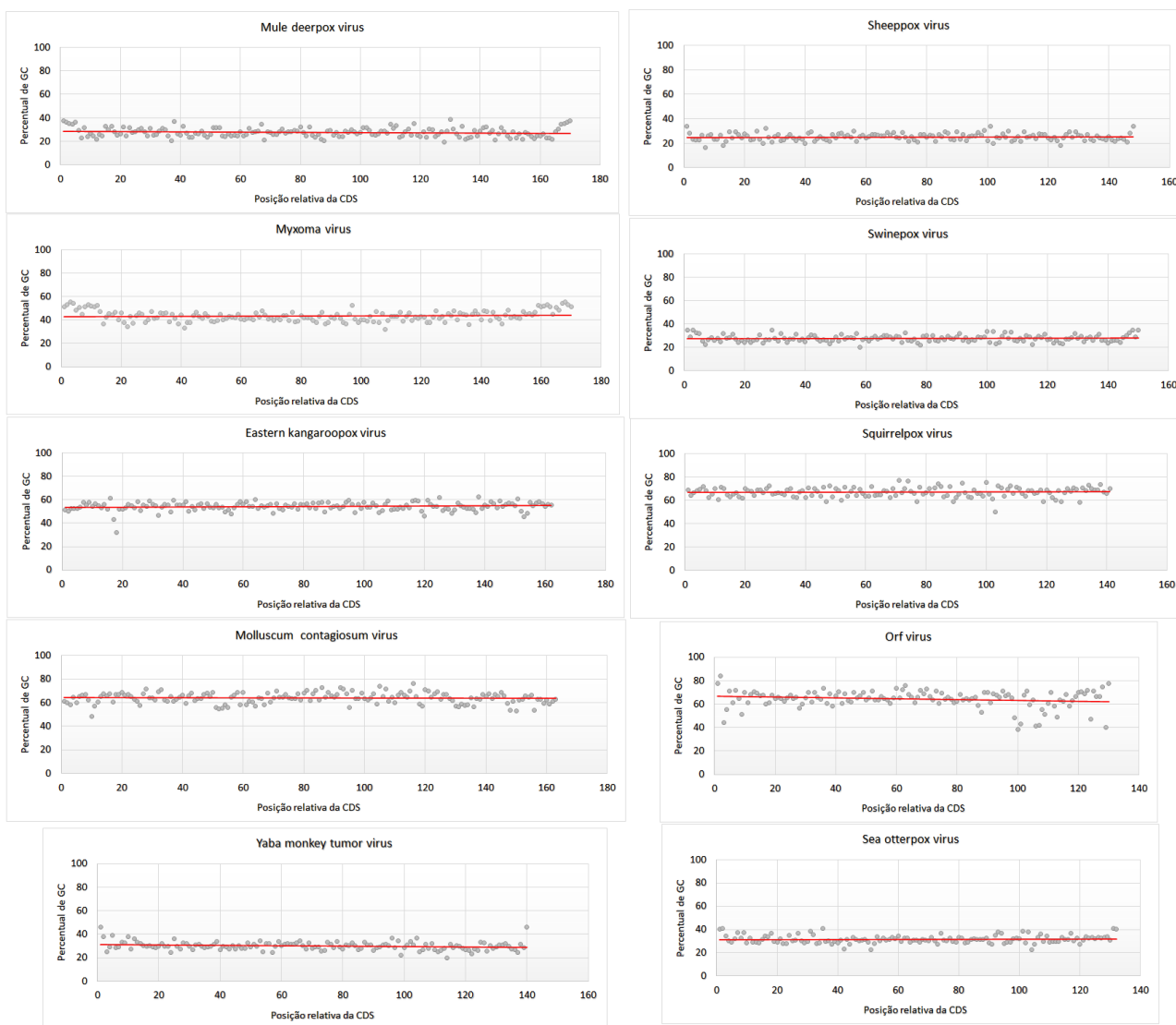


Figura 13. Representação da dispersão dos valores de conteúdo GC%, gene a gene, em genomas de Poxviridae (painel B). Representação gráfica da dispersão dos valores de percentual GC (identificados no eixo Y) para cada CDS anotada disponível (identificadas pelos pontos posicionados na extensão do eixo X), para os genomas de mule deerpox vírus, sheeppox vírus, myxoma vírus, swinepox vírus, eastern kangaroopox vírus, squirrelepox vírus, molluscum contagiosum vírus, orf vírus, yaba monkey tumor vírus

e sea otterpox vírus. Linha de tendência dos valores médios representada em vermelho, sobre os valores pontuais de GC% de cada genoma.

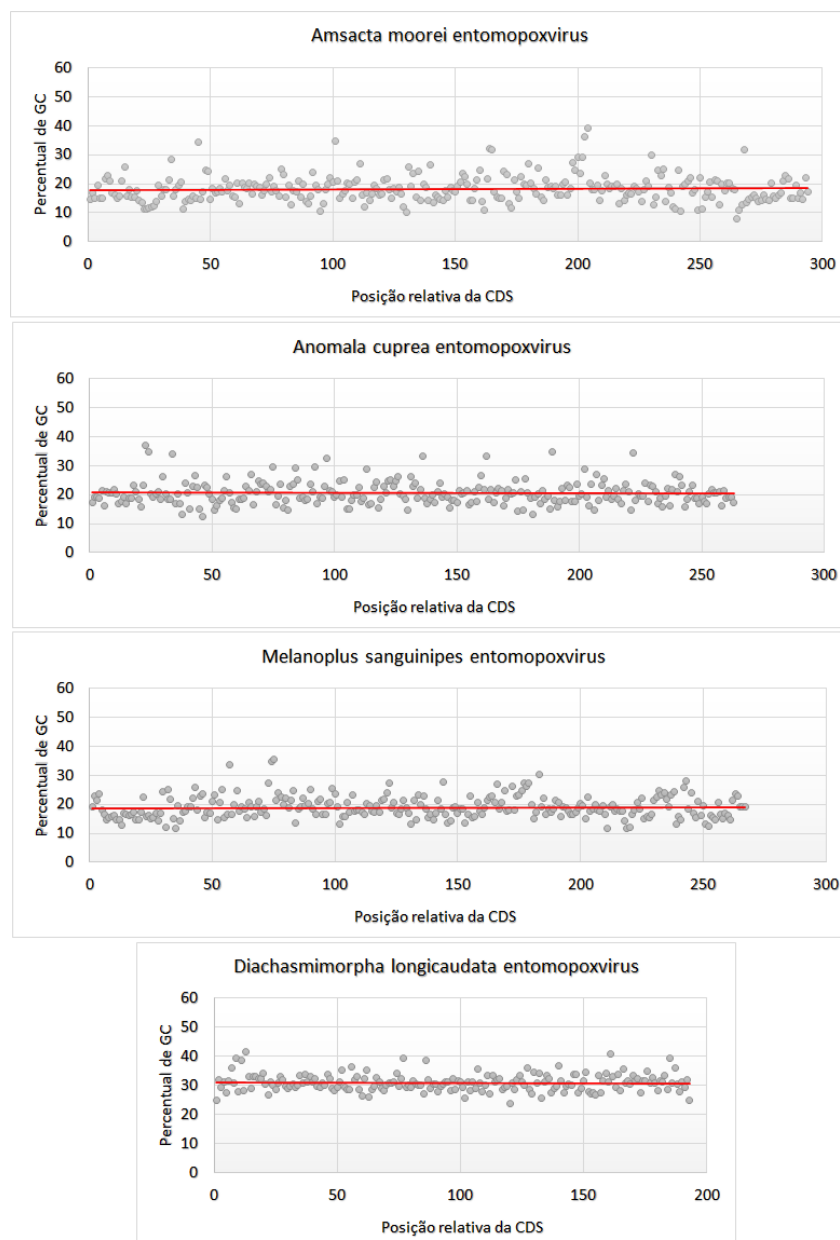


Figura 14. Representação da dispersão dos valores de conteúdo GC%, gene a gene, em genomas de Entomopoxvirinae. Representação gráfica da dispersão dos valores de percentual GC (identificados no eixo Y) para cada CDS anotada disponível

(identificadas pelos pontos posicionados na extensão do eixo X), para os genomas de *Amsacta moorei* entomopoxvirus, *Anomala cúprea* entomopoxvirus, *Melanoplus sanguinipes* entomopoxvirus e *Diachasmimorpha longicaudata* entomopoxvirus. Linha de tendência dos valores médios representada em vermelho, sobre os valores pontuais de GC% de cada genoma.

Em “Pithoviridae” (**figura 15**), com somente três genomas utilizados neste trabalho, sendo estes orpheovirus – com o maior número de regiões codificadoras – cedratvirus A11 e pithovirus sibericum – com o menor número de regiões codificadoras.

Observamos que a tendência de média dos valores GC% para cedratvirus A11 e pithovirus sibericum são de aproximadamente 40%, enquanto o valor médio observado em orpheovirus é menor, abaixo de 30%.

Em relação à dispersão dos valores GC% atribuídos a regiões codificadoras, observamos que o genoma de cedratvirus A11 apresenta menor amplitude, com valores majoritariamente concentrados próximos ao valor médio, enquanto nos demais genomas é possível observar maior amplitude – principalmente em orpheovirus.

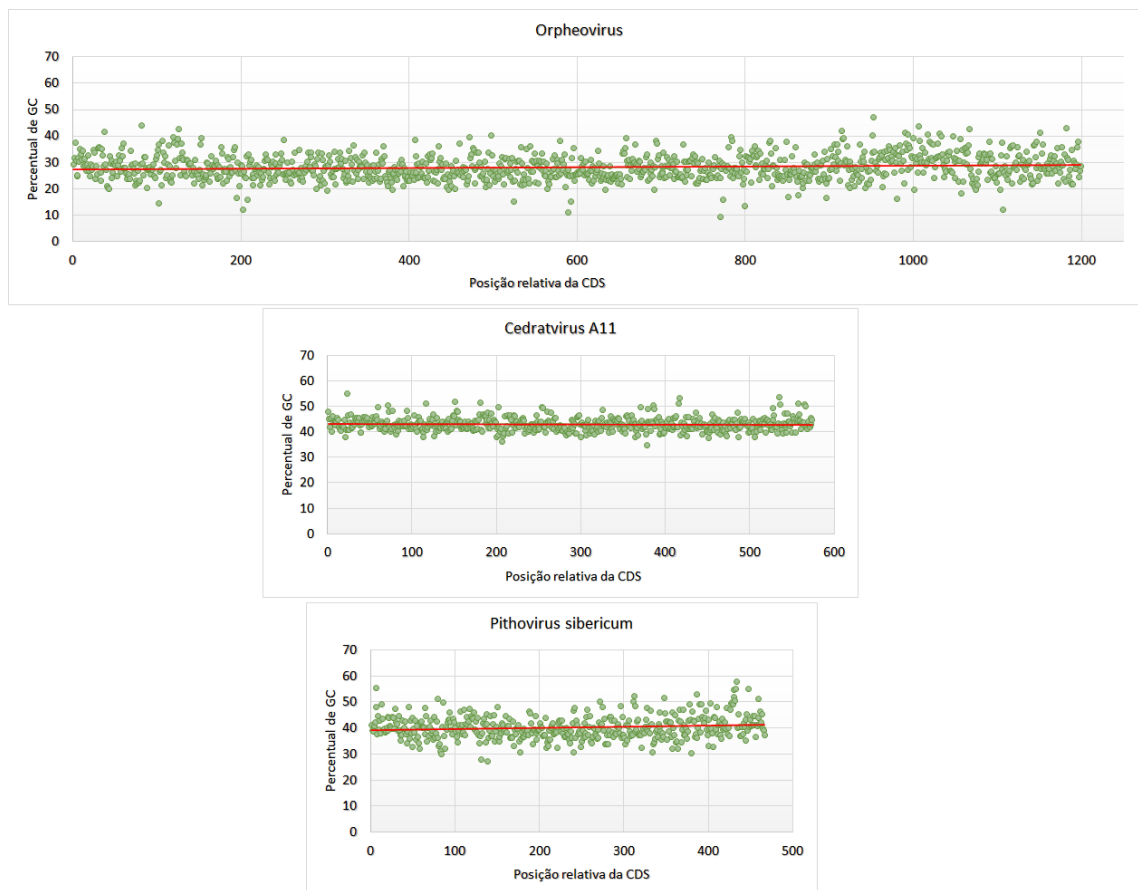


Figura 15. Representação da dispersão dos valores de conteúdo GC%, gene a gene, em genomas de "Pithoviridae". Representação gráfica da dispersão dos valores de percentual GC (identificados no eixo Y) para cada CDS anotada disponível (identificadas pelos pontos posicionados na extensão do eixo X), para os genomas de orpheovirus, cedratvirus A11 e pithovirus sibericum. Linha de tendência dos valores médios representada em vermelho, sobre os valores pontuais de GC% de cada genoma. Neste painel, orpheovirus, um pithoviridae-like está representado junto a "Pithoviridae".

6.3 Comparação da variação de GC% gene a gene intra-grupos

Com o intuito de avaliar comparativamente o GC% gene a gene de diferentes vírus no contexto de suas respectivas famílias, estabelecemos uma abordagem de análise que considerou não só a média de GC% para cada genoma, mas também a média geral de GC% de todos os membros analisados dentro de cada grupo taxonômico. Nesta representação, cada coluna representada equivale ao genoma de um vírus pertencente ao grupo taxonômico estudado. Cada linha colorida ao longo do genoma é a representação de uma CDS do mesmo.

Nos três gráficos, valores percentuais de GC baixos recebem cor azul e valores percentuais de GC alto recebem cor vermelha, de acordo com a proporção dos valores (ou seja, quanto menor o valor, mais intensa o tom de azul e, quanto maior o valor, mais intenso o tom de vermelho). *A cor amarela representa os valores médios. Em A, cada genoma é colorido em relação ao valor médio do próprio genoma, individualmente. Em B e C, os genomas são coloridos em uma mesma escala, agrupados, levando em consideração o valor médio de CDS do grupo taxonômico (B) ou o valor médio total do grupo taxonômico (C).* Para a melhor visualização possível das imagens geradas, considerando os genomas especialmente maiores, as próximas figuras estarão condicionadas de modo a ocupar o maior espaço possível dentro das margens do documento.

Nos mapas gerados para os genomas representados de membros da família *Phycodnaviridae* (**figura 16**), observamos no item A, inicialmente, valores de GC% constantes e próximos das respectivas médias individuais, com exceção de regiões contendo valores contrastantes. Em *Paramecium bursaria* *Chlorella* vírus 1, apresenta-se uma região disposta à segunda metade do genoma, onde evidencia-se valores significativamente abaixo e acima da média do genoma, sem sequência. *Emiliania huxleyi* vírus 86 apresenta uma região especialmente marcada, também disposta na segunda parte do genoma, onde encontra-se um *cluster* de genes com maior valor GC%. Outros *outliers* evidenciam-se nos quatro genomas analisados.

Em relação aos mapas referentes às classificações de acordo com GC% médio das CDS e GC% médio total dos genomas, não há diferença significativa – uma vez que as médias, respectivamente 41,06% e 40,6%, são muito próximas.

Apesar disto, tanto em B quando em C, observa-se que as médias de *Paramecium bursaria* *Chlorella* vírus 1 e *Emiliana huxleyi* vírus 86 se aproximam das médias em *Phycodnaviridae*, enquanto *Ectocarpus siliculosus* vírus 1 e *Heterosigma akashiwo* vírus 1 apresentam médias significativamente maior e menor, respectivamente.

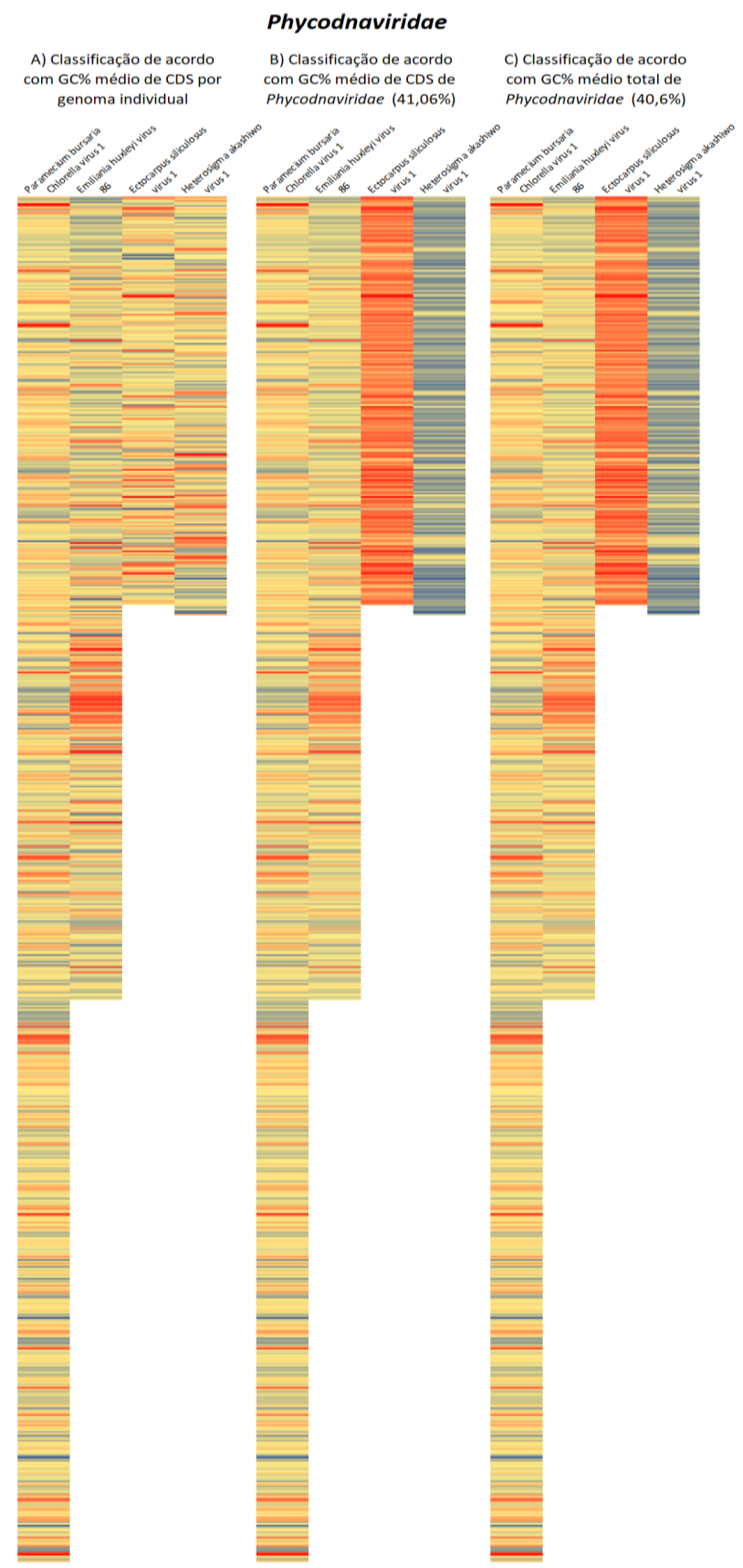


Figura 16. Mapas de classificação do valor de conteúdo GC% por cores, gene a gene, individualmente para cada genoma e intra-grupos taxonômicos. - *Phycodnaviridae*. Nesta representação gráfica, observamos nos itens A, B e C, respectivamente, classificações dos valores GC% de cada CDS em relação ao valor médio do genoma individual, valor médio das CDS observadas no grupo e valor médio genômico total (considerando regiões codificadoras e não codificadoras) do grupo. Nesta figura, estão representados os quatro genomas de *Phycodnaviridae* utilizados no trabalho.

Em *Mimiviridae* (**figura 17**), no mapa de distribuição dos valores GC% de acordo com as médias individuais, observa-se que os genomas apresentam valores de GC% constantes e próximos das respectivas médias. Apesar disto, são observadas pequenas regiões com agrupamentos de *outliers*, como no caso de Bodo saltans vírus, que apresentam um *cluster* de valores abaixo de sua média de GC% na extremidade final do genoma, ou Chrysochromulina ericina virus, que apresenta um *cluster* de valores acima de sua média de GC% na porção final do genoma.

Observando os mapas dos itens B e C, nos quais utilizam-se as médias de GC% relativa a regiões codificadoras do grupo (28,9%) e GC% relativa ao GC% total dos genomas do grupo (28,2%), não se observa diferença significativa. Os genomas de Acanthamoeba polyphaga mimivirus, samba vírus, tupanvirus soda lake, tupanvirus deep ocean e Aureococcus anophagefferens vírus apresentam-se mais próximos das médias de *Mimiviridae*, coloridos, em sua maior parte, de amarelo.

Os genomas de Cafeteria roenbergensis vírus, megavirus chiliensis, moomouvirus australiensis, Bodo saltans vírus, Chrysochromulina ericina vírus e cottonvirus japonicus apresentaram maior coloração azul, representando, portanto, valores menores em relação às médias de GC%, enquanto Tetraselmis viridis vírus apresentou maior coloração avermelhada, representando um genoma de valores GC% maiores do que as médias. O genoma de Phaeocystis globosa vírus group 1, embora classificado, majoritariamente, de amarelo, apresentou uma notável quantidade de CDS de valor GC% elevado.

Mimiviridae

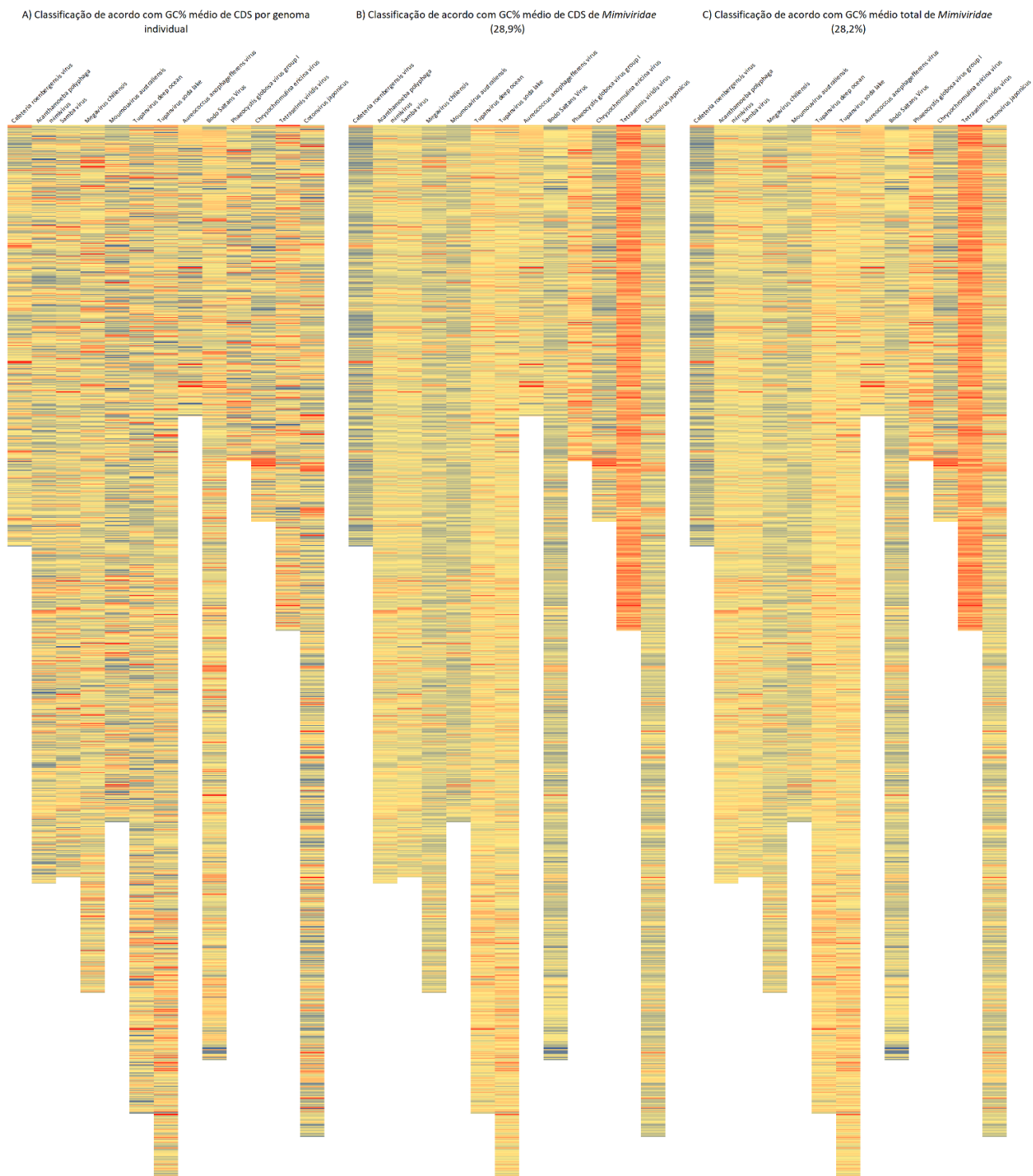


Figura 17. Mapas de classificação do valor de conteúdo GC% por cores, gene a gene, individualmente para cada genoma e intra-grupos taxonômicos. -

Mimiviridae. Nesta representação gráfica, observamos nos itens A, B e C, respectivamente, classificações dos valores GC% de cada CDS em relação ao valor médio do genoma individual, valor médio das CDS observadas no grupo e valor médio genômico total (considerando regiões codificadoras e não codificadoras) do grupo. Nesta figura, estão representados os 13 genomas de *Mimiviridae* utilizados no trabalho.

Em *Iridoviridae* (**figura 18**), são observados gráficos de classificação de GC% de acordo com a média individual nos quais os *outliers* estão mais evidentes – em relação aos grupos taxonômicos previamente descritos, *Phycodnaviridae* e *Mimiviridae*. Os genomas de invertebrate iridescente vírus 3 e invertebrate iridescente vírus 6 apresentam esta característica de forma mais pronunciada, principalmente em comparação a lymphocystis Disease vírus 1.

Em relação aos mapas B e C, considerando as médias de GC% relativa a regiões codificadoras do grupo (40,3%) e GC% relativa ao GC% total dos genomas do grupo (39%), não se observa diferença significativa. O genoma de Decapod iridescente vírus 1 apresenta uma notável quantidade de CDS classificadas como outliers, considerando ambas as médias de *Iridoviridae*, das quais a maioria apresentou valores GC% abaixo da média.

Ainda, os genomas de frog virus 3 e invertebrate iridescente vírus 3 apresentam-se coloridos quase inteiramente de vermelho intenso, reforçando que estes vírus possuem CDS de GC% médio consideravelmente acima da média do grupo taxonômico.

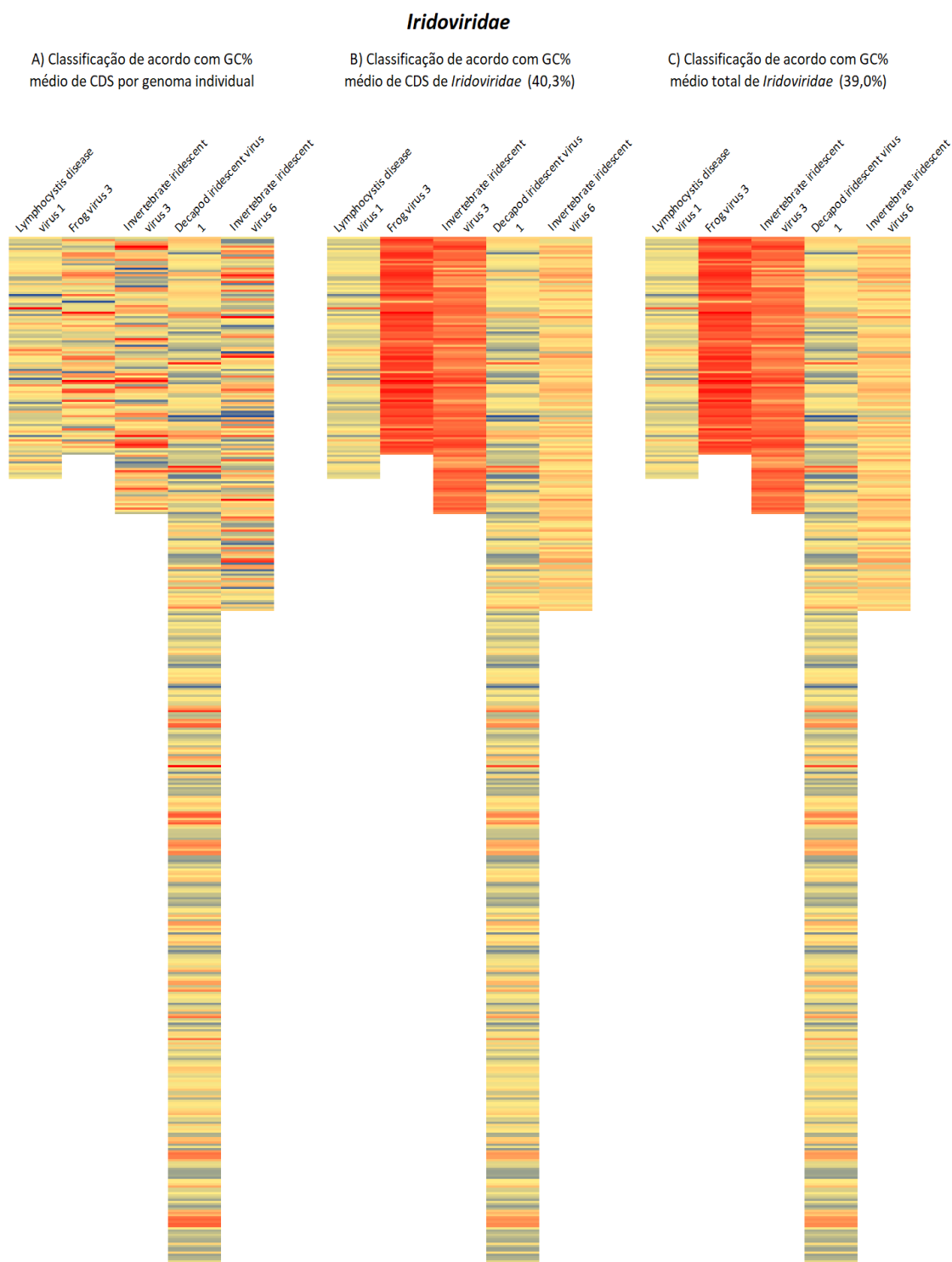


Figura 18. Mapas de classificação do valor de conteúdo GC% por cores, gene a gene, individualmente para cada genoma e intra-grupos taxonômicos. - *Iridoviridae*. Nesta representação gráfica, observamos nos itens A, B e C,

respectivamente, classificações dos valores GC% de cada CDS em relação ao valor médio do genoma individual, valor médio das CDS observadas no grupo e valor médio genômico total (considerando regiões codificadoras e não codificadoras) do grupo. Nesta figura, estão representados os cinco genomas de *Iridoviridae* utilizados no trabalho.

A análise do táxon *Marseilleviridae* (**figura 19**) apresentou uma notável consistência em relação às classificações por média de GC% das CDS por genoma, GC% médio das CDS do grupo (43,8%) e GC% médio total do grupo (43,4%). De modo geral, os cinco genomas trabalhados, apresentam um padrão conservado de distribuição de valor GC% associado a regiões codificadoras, próximo da média individual e do grupo taxonômico, com quantidade semelhante de *outliers*.

Embora os cinco genomas tenham apresentado certa consistência, observa-se que o genoma de golden marseillevírus possui CDS *outliers* um pouco mais pronunciadas quando considerando as médias individuais dos genomas. Nos mapas B e C, no entanto, esta diferença é atenuada.

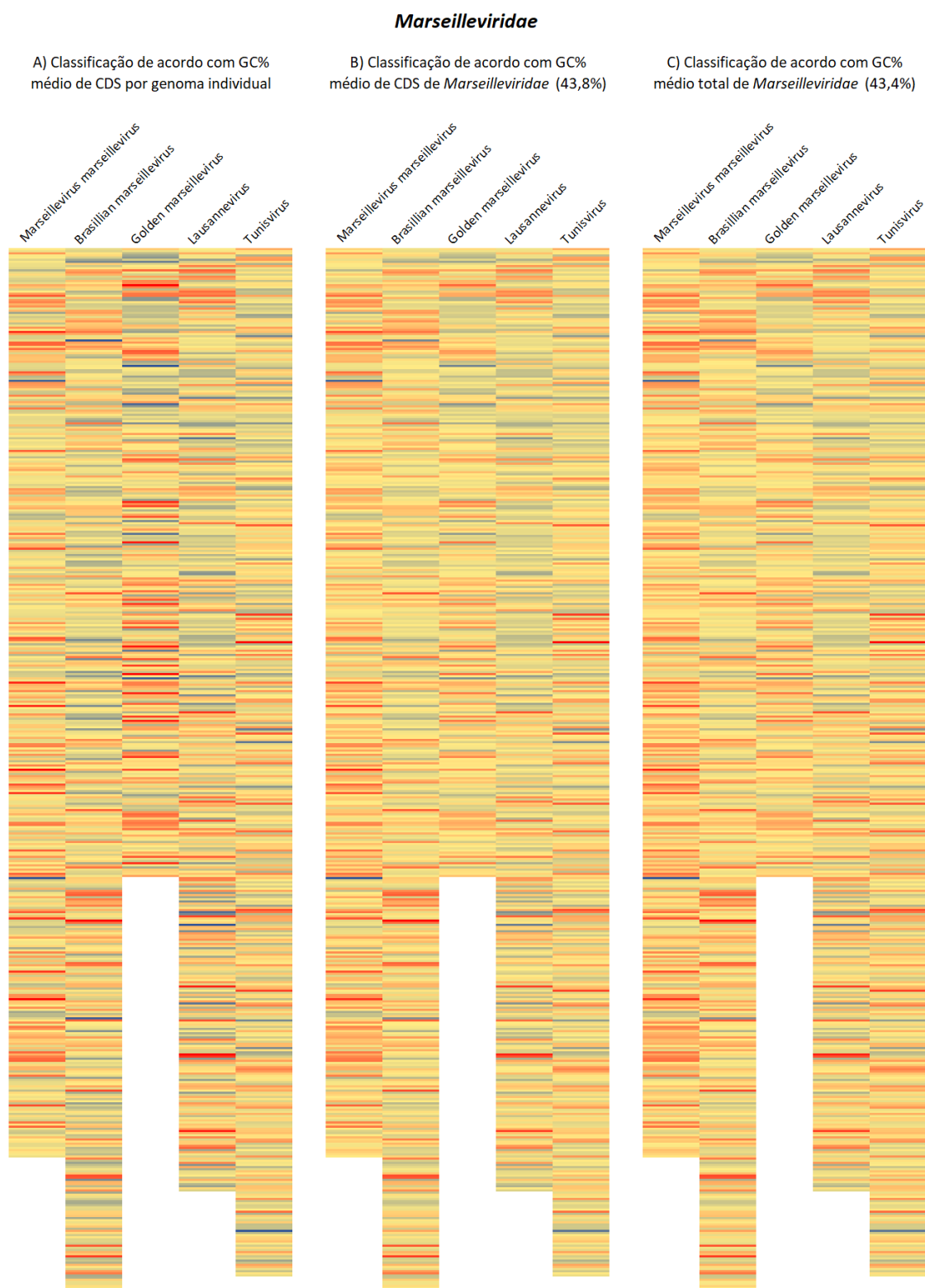


Figura 19. Mapas de classificação do valor de conteúdo GC% por cores, gene a gene, individualmente para cada genoma e intra-grupos taxonômicos. -

Marseilleviridae. Nesta representação gráfica, observamos nos itens A, B e C, respectivamente, classificações dos valores GC% de cada CDS em relação ao valor médio do genoma individual, valor médio das CDS observadas no grupo e valor médio genômico total (considerando regiões codificadoras e não codificadoras) do grupo. Nesta figura, estão representados os cinco genomas de *Marseilleviridae* utilizados no trabalho.

Em *Asfarviridae* (**figura 20**), em relação à classificação de acordo com as médias de GC% das CDS de cada genoma, não é possível identificar nenhum padrão específico. Apesar disto, se observa que em african swine fever vírus e kaumoebavírus regiões com elevada quantidade de *outliers* se localizam nas extremidades dos genomas.

Nos mapas de caracterização de valores GC% de acordo com a média de CDS e média genômica, 38,3% e 38,1%, respectivamente (B e C), não se observa diferença significativa, uma vez que os valores médios são muito próximos. Entretanto, em comparação à classificação individual de A, nos mapas B e C é possível identificar o genoma de faustovírus e12 como mais próximo do valor médio de GC% do grupo e o genoma de african swine fever vírus como o mais representativo do grupo, uma vez que praticamente não sofre alteração nas três classificações.

Os genomas de pacmanvírus S19 e kaumoebavírus, por outro lado, apresentam cores mais azuladas e avermelhadas, respectivamente, indicando genomas com mais regiões codificadoras com valor GC% abaixo e acima da média de *Asfarviridae*.

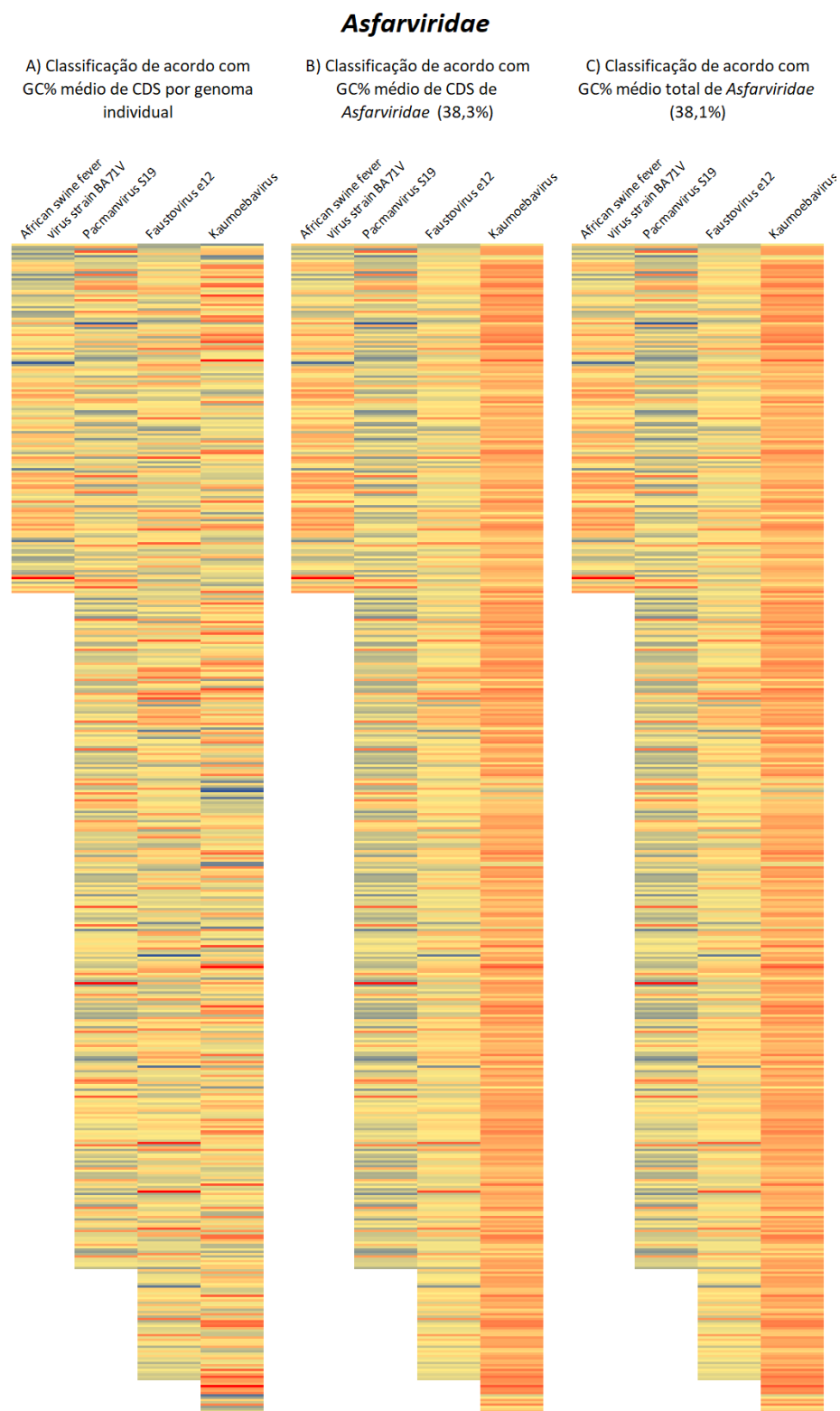


Figura 20. Mapas de classificação do valor de conteúdo GC% por cores, gene a gene, individualmente para cada genoma e intra-grupos taxonômicos. -

***Asfarviridae*.** Nesta representação gráfica, observamos nos itens A, B e C, respectivamente, classificações dos valores GC% de cada CDS em relação ao valor médio do genoma individual, valor médio das CDS observadas no grupo e valor médio genômico total (considerando regiões codificadoras e não codificadoras) do grupo. Nesta figura, estão representados os quatro genomas de *Asfarviridae* utilizados no trabalho.

A análise de membros da família *Poxviridae* (**figura 21**), de acordo com a classificação dos valores GC% em relação às médias individuais dos genomas (A), observa-se um padrão na classificação das regiões codificadoras das extremidades dos genomas. Isto é, os genomas dos poxvírus, em sua maioria, apresentam valores GC% atribuídos às regiões codificadoras invertidos e espelhados em suas extremidades. Este padrão é mais facilmente visualizado nos genomas onde os valores GC% das extremidades são considerados outliers, como é o caso de: fowlpoxvirus, mule deerpox vírus, myxoma vírus, sea otterpoxvirus, vaccínia vírus, Cotia vírus, salmon gillpox vírus, orf vírus, swinepox vírus e *Melanoplus sanguinipes* entomopoxvirus.

Ainda, considerando as regiões descritas acima, observou-se que os valores de outliers localizados nestas regiões são, em sua maioria, significativamente maiores do que os valores GC% médios de seus respectivos genomas. Esta característica está, especialmente, evidente em myxoma vírus, fowlpox vírus, mule deerpoxvirus e vaccinia vírus.

Além destas regiões de notável variação de valores GC%, outras regiões contendo grande quantidade de outliers foram identificadas. De modo geral, todos os genomas de poxvírus representados neste trabalho possuem variações do perfil GC% de suas CDS mais destacadas do que as observadas nos demais grupos de *Nucleocyotoviricota*.

Comparando as representações gráficas que utilizam o valor médio das CDS e o valor médio genômico dos poxvírus (B e C), com respectivos 36,3% e 35,9%, não se observa diferença significativa. No entanto, em ambas as representações, é possível visualizar genomas que apresentam um elevado número de CDS com valor GC% abaixo

da média do grupo (em azul), sendo os mais significativos os genomas de *Amsacta moorei* entomopoxvirus, *Melanoplus sanguinipes* entomopoxvirus, *Anomala cúprea* entomopoxvirus e *Cotia* vírus e, também, é possível observar genomas que apresentam elevado número de CDS com valor GC% acima da média do grupo (em vermelho), como *squirrelepox* vírus, *orf* vírus, *nile crocodilepox* vírus e *molluscum contagiosum* vírus.

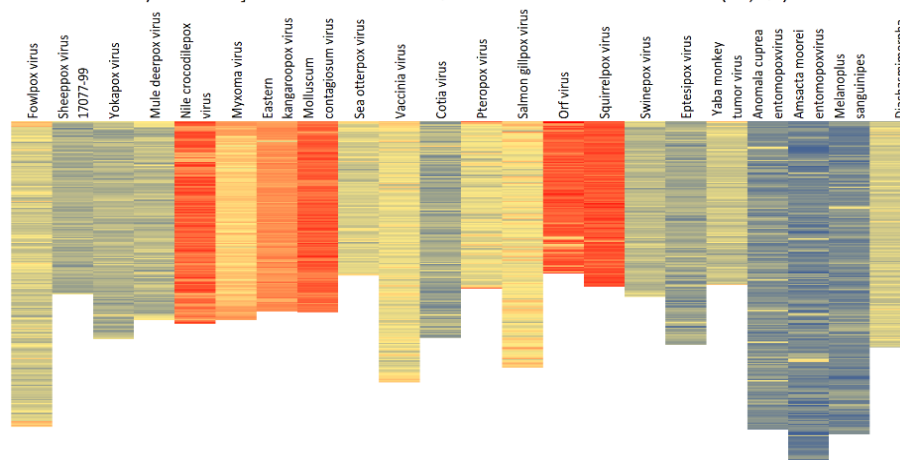
Ainda, nestas representações que utilizam o valor médio de GC% do grupo, a característica observada no item A ainda pode ser visualizada com clareza para alguns vírus, sendo alguns destes: *fowlpox* virus, *mule deerpox* vírus, *myxoma* vírus, *pteropox* vírus e *swinepox* vírus.

Poxviridae

A) Classificação de acordo com GC% médio de CDS por genoma individual



B) Classificação de acordo com GC% médio de CDS de *Poxviridae* (36,3%)



C) Classificação de acordo com GC% médio de CDS de *Poxviridae* (35,9%)

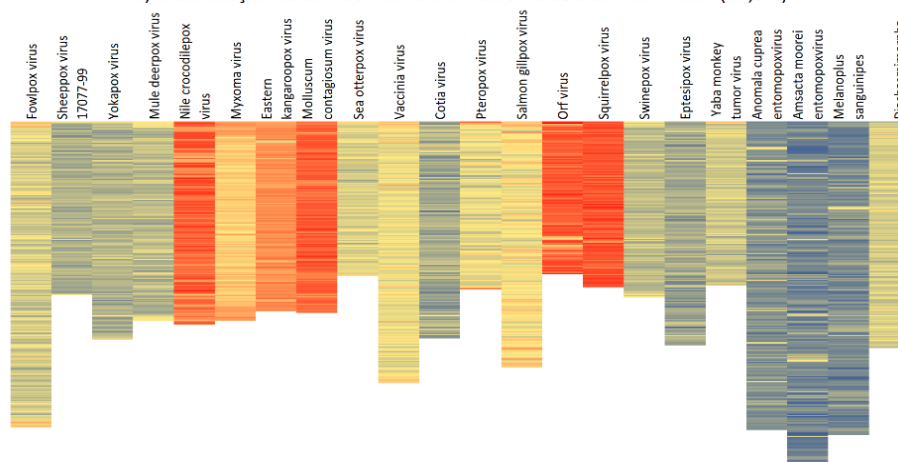


Figura 21. Mapas de classificação do valor de conteúdo GC% por cores, gene a gene, individualmente para cada genoma e intra-grupos taxonômicos. -

Poxviridae. Nesta representação gráfica, observamos nos itens A, B e C, respectivamente, classificações dos valores GC% de cada CDS em relação ao valor médio do genoma individual, valor médio das CDS observadas no grupo e valor médio genômico total (considerando regiões codificadoras e não codificadoras) do grupo. Nesta figura, estão representados os 22 genomas de *Poxviridae* utilizados no trabalho, incluindo os genomas pertencentes à subfamília *Entomopoxvirinae*.

Nos mapas de classificação dos valores GC% das regiões codificadoras de “Pithoviridae” (**figura 22**) em relação às médias individuais dos genomas estudados, observa-se que pithovirus sibericum apresenta regiões de CDS outliers em evidência, tanto para valores abaixo quanto acima de sua média. O genoma de orpheovírus apresenta em sua porção final uma concentração de outliers de valor acima de sua média, embora bem espaçados.

Já ao comparar as representações gráficas que utilizam o valor médio das CDS e o valor médio genômico do grupo (B e C), com respectivos 37,1% e 34,5%, evidencia-se a diferença entre a tendência de valores GC% individual dos três vírus: pithovírus sibericum aproxima-se mais da média do grupo, apesar de apresentar muitos valores GC% mais elevados; cedratvírus A11 apresenta perfil GC% de CDS majoritariamente mais elevado e orpheovírus apresenta perfil GC% de CDS mais baixo, ambos em relação à média.

Ainda, é possível perceber uma diferença sutil na intensidade das cores entre os mapas B e C, apesar de similares, o que indica que para o *dataset* de “Pithoviridae” analisado existe uma diferença potencialmente significativa entre o perfil GC% das regiões codificadoras e o perfil GC% dos genomas em sua totalidade, incluindo regiões intergênicas.

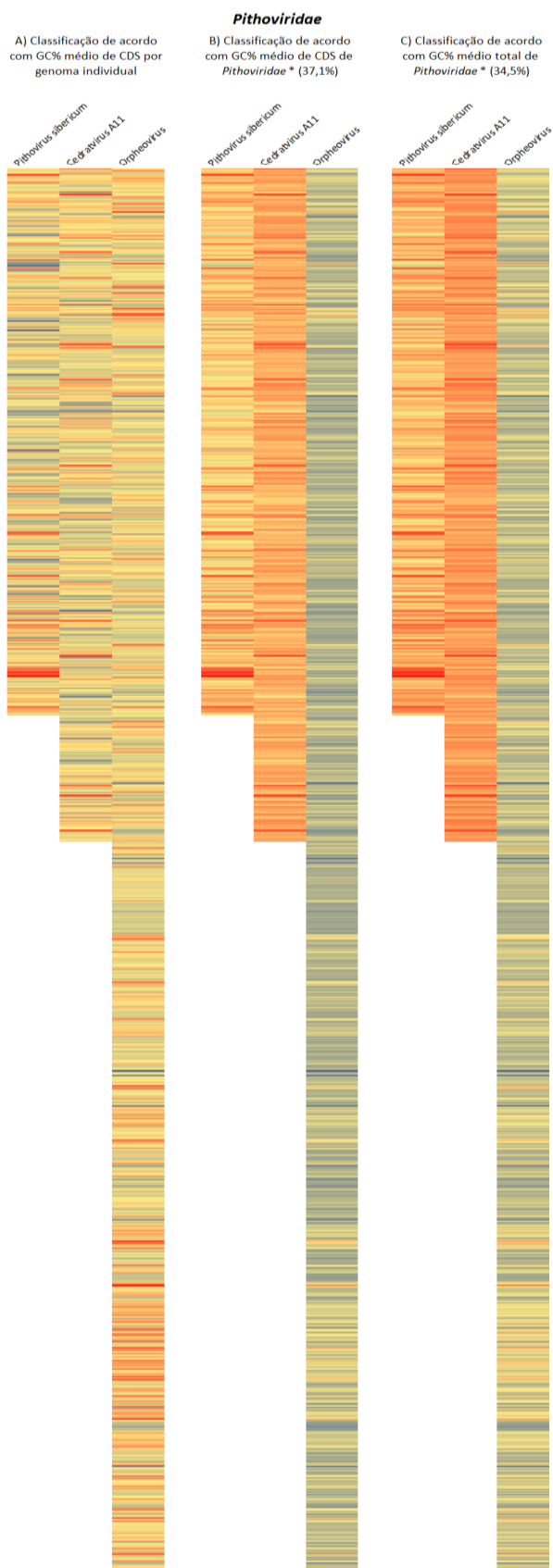


Figura 22. Mapas de classificação do valor de conteúdo GC% por cores, gene a gene, individualmente para cada genoma e intra-grupos taxonômicos. - *Pithoviridae*. Nesta representação gráfica, observamos nos itens A, B e C, respectivamente, classificações dos valores GC% de cada CDS em relação ao valor médio do genoma individual, valor médio das CDS observadas no grupo e valor médio genômico total (considerando regiões codificadoras e não codificadoras) do grupo. Nesta figura, estão representados os três genomas de “Pithoviridae” utilizados no trabalho.

A classificação por cor dos valores GC% atribuídos às CDS dos 61 genomas de *Nucleocytoviricota* (**figura 23**) – em uma escala onde o valor médio de GC% genômico dos 61 genomas é colorido em amarelo, o maior valor de GC% observado para uma região codificadora é colorido em vermelho e o menor valor de GC% observado para uma região codificadora é colorido em azul – ilustra, de modo geral, qual o padrão (se há algum) de distribuição de valores GC% em cada família.

Phycodnaviridae, *Marseilleviridae*, *Asfarviridae* e “*Pithoviridae*” são as famílias viriais que possuem os integrantes cuja distribuição dos valores GC% mais se assemelha do valor médio GC% para NCLDV – embora apresentem exceções, como *Ectocarpus siliculosus* vírus 1 e orpheovírus.

Os genomas de representantes da família *Mimiviridae*, de modo geral, classificam-se como portadores de CDS com valores GC% consideravelmente abaixo da média de *Nucleocytoviricota*, com a exceção de *Tetraselmis viridis* vírus, que se aproxima da distribuição média do grupo. Apesar disto, os genomas com maior representação de valores GC% baixos não estão em *Mimiviridae*.

Iridoviridae, nesta classificação, possui genomas com uma representação de valores baixos, *Lymphocystis Disease* vírus 1 e decapod iridescente vírus 3, e genomas com representação de valores elevados, frog vírus 3 e invertebrate iridescente vírus 3.

A maior variação observada em membros de *Nucleocytoviricota* nesta classificação está em membros da família *Poxviridae*, que apresenta os maiores e menores valores GC% representados em genomas de sua família. Destes, podemos destacar os genomas que apresentam majoritariamente valores GC% baixos, e os mais baixos observados em *Nucleocytoviricota*, sendo estes pertencentes à subfamília *Entomopoxvirinae*: *Anomala cúprea* entomopoxvírus, *Amsacta moorei* entomopoxvírus e *Melanoplus sanguinipes* entomopoxvírus; e, também, podemos destacar os genomas que apresentam majoritariamente valores GC% elevados: *nile crocodilepox* vírus, *molluscum contatiosum* vírus, *orf* vírus e *swinepox* vírus.

Ainda, genomas que não foram trabalhados nas seções anteriores – *Spodoptera frugiperda* ascovirus 1A, *yaravírus* brasiliense, *pandoravírus quercus*, *medusavírus* e

mollivírus sibericum – e estão incluídos neste mapa de classificação dos valores GC% de *Nucleocytoviricota* apresentaram, coincidentemente, genomas cujas regiões codificadoras apresentavam em sua maioria valores GC% significativamente maiores em relação à média do filo.

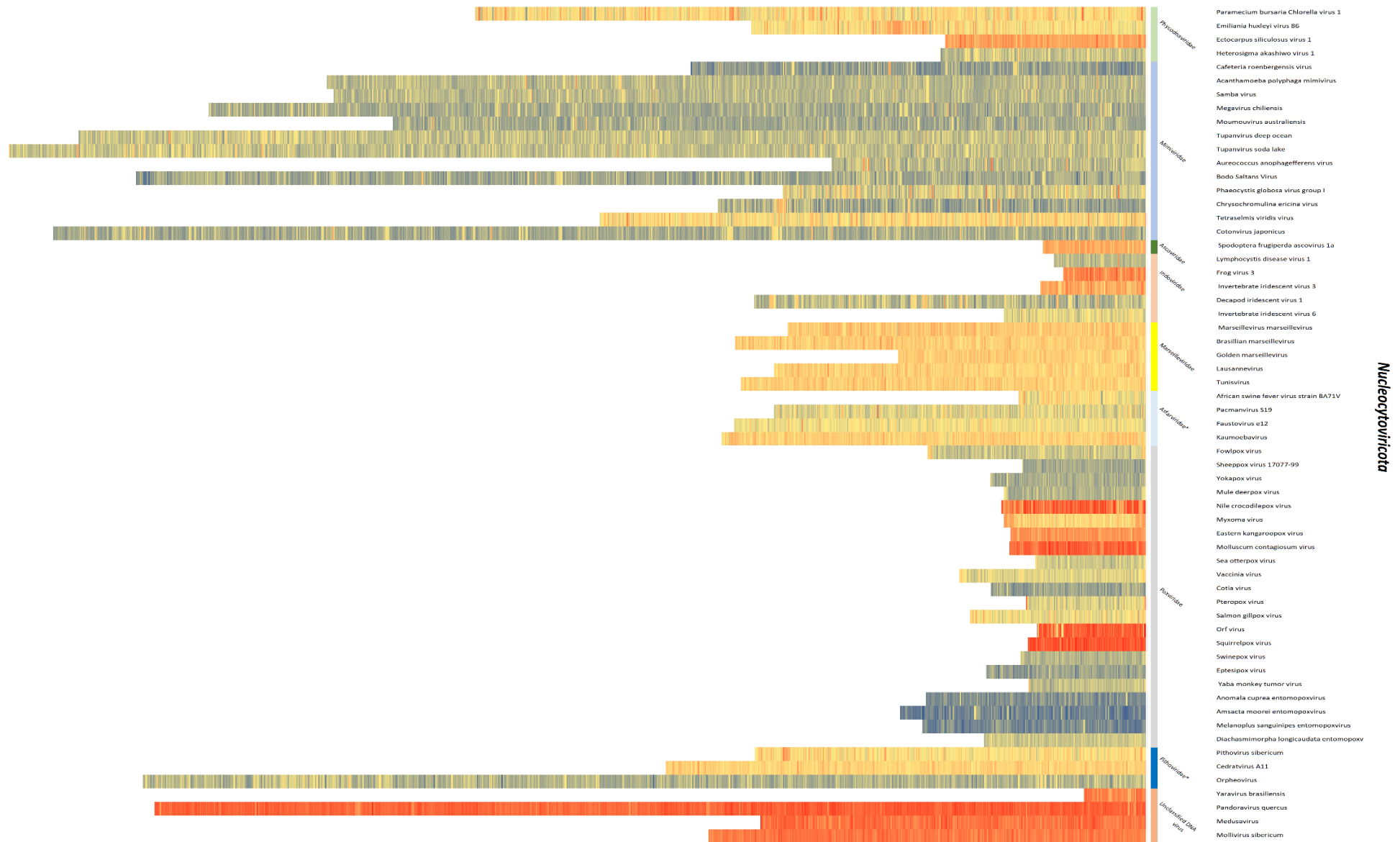


Figura 23. Mapas de classificação do valor de conteúdo GC% por cores, gene a gene, individualmente para cada genoma e intra-grupos taxonômicos. - *Nucleocytoviricota*. Nesta representação gráfica, observamos nos itens A, B e C, respectivamente, classificações dos valores GC% de cada CDS em relação ao valor médio do genoma individual, valor médio das CDS observadas no

grupo e valor médio genômico total (considerando regiões codificadoras e não codificadoras) do grupo. Nesta figura, estão representados todos os 61 genomas de *Nucleocytoviricota* incluídos neste trabalho.

6.4 Regiões codificadoras de percentagens GC% máximas e mínimas

Apresentam-se nesta seção tabelas que compilam informações disponíveis no NCBI acerca das CDS de maiores e menores valores GC% (três maiores e três menores) em cada genoma estudado, de acordo com o grupo taxonômico correspondente. Estão representados nas tabelas o nome do vírus, o valor percentual de GC da CDS, o número de identificação da proteína predita e a descrição disponível no banco de dados público.

As regiões codificadoras de maior e menor valor GC% observadas em *Phycodnaviridae* são de, respectivamente, *Paramecium bursaria* *Chlorella* vírus 1 com 64,63% e *Heterosigma akashiwo* vírus 1 com 18,61%. Ambas as CDS anotadas são descritas como proteínas hipotéticas. A maioria das CDS outliers em *Phycodnaviridae* apresentaram descrições para proteínas hipotéticas. Nenhuma tendência foi observada no grupo.

Tabela 2. Valores GC% máximos e mínimos observados em *Phycodnaviridae*.

Phycodnaviridae - CDS máximas e mínimas

Vírus	GC%	ID da proteína predita	Descrição
<i>Paramecium bursaria</i> <i>Chlorella</i> virus 1	64,36	NP_048416.1	Proteína hipotética
	64,63	YP_004678872.1	Proteína hipotética
	64,63	YP_004679009.1	Proteína hipotética
	20,27	NP_048936.1	Proteína hipotética
	20,63	YP_004678995.1	Proteína hipotética
	23,19	YP_004679000.1	Proteína hipotética
<i>Emiliania huxleyi</i> virus 86	57,74	YP_294081.1	Proteína hipotética
	59,06	YP_293958.1	Proteína hipotética - CDS rica em prolina, similar a " <i>Volvox carteri nagariensis pherophorin-DZ1 protein precursor</i> "
	60,05	YP_294122.1	Proteína hipotética - CDS rica em prolina, similar a " <i>Chlamydomonas incerta SAG1 plus agglutinin</i> "
	25	YP_293991.1	Proteína hipotética
	27,45	YP_294012.1	Proteína hipotética
	29,69	YP_294117.1	Proteína hipotética
<i>Ectocarpus siliculosus</i> virus 1	63,6	NP_077699.1	EsV-1-214
	63,82	NP_597948.1	Proteína hipotética
	64,16	NP_077542.1	Proteína colágeno-like 1
	37,25	NP_077662.1	EsV-1-177
	38,81	NP_077520.1	EsV-1-35
	39,15	NP_077518.1	EsV-1-33
<i>Heterosigma akashiwo</i> virus 1	47,18	YP_009507549.1	Proteína hipotética
	43,39	YP_009507598.1	Proteína de membrana externa putativa
	42,29	YP_009507610.1	Proteína hipotética
	22	YP_009507410.1	Proteína hipotética
	20,5	YP_009507642.1	Proteína hipotética
	18,61	YP_009507622.1	Proteína hipotética

As regiões codificadoras de maior e menor valor GC% observadas em *Mimiviridae* (**tabela 3**) são de, respectivamente, Chrysochromulina ericina virus com 62,55% e Cafeteria roenbergensis virus com 9,68%. Destas, a CDS de maior valor GC% apresenta a descrição de “proteína contendo domínio HGM *box*”, enquanto a CDS de menor valor GC% é descrita como proteína hipotética.

A maioria das CDS outliers em membros da família *Mimiviridae* apresentaram descrições para proteínas hipotéticas. Nenhuma tendência foi observada no grupo, mas foram identificadas proteínas de capsídeo e proteína de membrana em Tetraselmis viridis vírus e cotonvírus japonicus, respectivamente.

Tabela 3. Valores GC% máximos e mínimos observados em *Mimiviridae*

Mimiviridae - CDS máximas e mínimas				Tupanvirus soda lake	49,95	QKU35887.1	Proteína hipotética
Vírus	GC%	ID da proteína predita	Descrição		50,23	QKU35673.1	Proteína contendo repetição e colágeno
					51,67	QKU35266.1	ORFan putativa
Cafeteria roenbergensis virus	17,12	QKU34920.1	ORFan putativa		17,62	QKU35449.1	ORFan putativa
	18,84	QKU34659.1	Proteína hipotética		56,78	YP_009052410.1	Proteína hipotética
	58,14	YP_009052387.1	Proteína hipotética	59,92	YP_009052263.1	Proteína hipotética	
	17,58	YP_009052438.1	Proteína hipotética	18,1	YP_009052314.1	Proteína hipotética	
	18,25	YP_009052391.1	Proteína hipotética	44,44	ATZ80928.1	Proteína hipotética	
	46,77	ATZ80793.1	Proteína hipotética	48,83	ATZ80929.1	Proteína hipotética	
Acanthamoeba polyphaga mimivirus	10	ATZ81255.1	Proteína hipotética	11,71	ATZ80146.1	Proteína hipotética	
	12,82	ATZ80137.1	Proteína hipotética	56,77	YP_008052601.1	Proteína contendo domínio SWIB	
	60,61	YP_008052352.1	Proteína hipotética	62,07	YP_008052510.1	Proteína hipotética	
	18,73	YP_008052647.1	Proteína hipotética	19,37	YP_008052575.1	Proteína hipotética	
	20,27	YP_008052498.1	Proteína hipotética	57,09	YP_009173431.1	Proteína contendo repetição e colágeno	
	57,11	YP_009173693.1	Proteína hipotética	62,55	YP_009173695.1	Proteína contendo domínio HMG box	
Samba virus	13,33	YP_009173413.1	Proteína hipotética	14,62	YP_009173311.1	Proteína hipotética	
	14,92	YP_009173455.1	Proteína hipotética	56,18	AUF82396.1	Proteína hipotética	
	57,77	AUF82093.1	Principal proteína de capsídeo	60,14	AUF82712.1	Proteína hipotética	
	23,19	AUF82479.1	Proteína hipotética	24,69	AUF82109.1	Proteína hipotética	
	25,45	AUF82588.1	Proteína hipotética	49,5	BCS83272.1	Proteína hipotética	
Megavirus chiliensis	50,14	BCS82517.1	Proteína hipotética	51,16	BCS83461.1	ORFan putativa	
	17,16	BCS82495.1	ORFan putativa	17,82	BCS83524.1	Proteína hipotética	
	17,95	BCS82583.1	Proteína de membrana				
Moumouvirus australiensis							
Tupanvirus deep ocean							

Em *Iridoviridae*, as regiões codificadoras de maior e menor valor GC% observadas são de, respectivamente, frog vírus 3 com 69,09% e invertebrate iridescente vírus 6 com 15,15%. Ambas as CDS anotadas são descritas como proteínas hipotéticas.

A maioria das CDS outliers em membros da família *Iridoviridae* apresentaram descrições para proteínas hipotéticas. Nenhuma tendência foi observada no grupo, embora tenham sido identificados um fator de replicação com putativa ação de ligação de DNA ou empacotamento em lymphocystis disease vírus 1, e uma DNA-ligase em decapod iridescente vírus 1.

Tabela 4. Valores GC% máximos e mínimos observados em Iridoviridae.

Iridoviridae - CDS máximas e mínimas

Vírus	GC%	ID da proteína predita	Descrição
Lymphocystis disease virus 1	39,09	NP_078716.1	Proteína hipotética
	40,62	NP_579903.1	Proteína hipotética
	53,53	NP_078660.1	Proteína tipo colágeno
	19,64	NP_078652.1	Proteína hipotética
	20,63	NP_078710.1	Proteína hipotética
	22,22	NP_078747.1	Fator de replicação e/ou proteína de ligação/empacotamento de DNA putativa
Frogvirus 3	65,86	YP_031653.1	Proteína hipotética
	68,18	YP_031613.1	Proteína hipotética
	69,09	YP_031644.1	Proteína hipotética
	36,6	YP_031608.1	Proteína hipotética
	40,98	YP_031640.1	Proteína hipotética
	45,16	YP_031665.1	Proteína hipotética
Invertebrate iridescent virus 3	59,68	YP_654663.1	Proteína hipotética
	60,82	YP_654638.1	Proteína hipotética
	60,86	YP_654577.1	Proteína hipotética
	36,6	YP_654587.1	Proteína hipotética
	38,34	YP_654622.1	Proteína hipotética
	38,63	YP_654595.1	Proteína hipotética
Decapod iridescent virus 1	45,57	ATE87064.1	DNA-ligase NAD dependente
	45,68	ATE87129.1	Proteína hipotética
	46,75	ATE87046.1	Proteína hipotética
	25,29	ATE87062.1	Proteína hipotética
	25,32	ATE87091.1	Proteína hipotética
	26,59	ATE87089.1	Proteína hipotética
Invertebrate iridescent virus 6	52,87	NP_149567.1	Proteína 104L
	53,62	NP_149520.1	Proteína 057L
	57,67	NP_149703.1	Proteína 240R
	15,15	NP_149544.1	Proteína 081L
	17,46	NP_149667.1	Proteína 204L
	18,3	NP_149546.1	Proteína 083L

As CDS consideradas *outliers* de maior valor em membros da família *Marseilleviridae* foram identificadas nos genomas de tunisvirus e brazilian marseillevirus, ambas com 57,36%, enquanto o menor valor foi identificado em marseillevirus marseillevirus, com 27,88%. As três CDS com maiores e menor valores observados em membros da família *Marseilleviridae* foram descritas como proteínas hipotéticas.

Não foi observado nenhum padrão em relação às proteínas codificadas pelas CDS *outliers* em valores GC% na família. Entretanto, foram identificadas proteínas histonas H2B/H2A de fusão em lausannevirus e brazilian marseillevirus, ambas com conteúdo GC% elevado.

Tabela 5. Valores GC% máximos e mínimos observados em *Marseilleviridae*.

***Marseilleviridae* - CDS máximas e mínimas**

Vírus	GC%	ID da proteína predita	Descrição
Marseillevirus marseillevirus	53,07	YP_003406971.1	Tioredoxina
	54,65	YP_003407109.1	Proteína hipotética
	54,71	YP_003407001.1	Proteína hipotética
	27,88	YP_003407052.1	Proteína hipotética
	30,19	YP_003406818.1	Proteína hipotética
	37,23	YP_003406786.1	Proteína hipotética
Brazilian marseillevirus	52,99	AMQ10945.1	Proteína de fusão histona H2B/H2A
	53,85	AMQ10983.1	Proteína hipotética
	57,36	AMQ10825.1	Proteína hipotética
	33,33	AMQ10552.1	Proteína hipotética
	34,02	AMQ10871.1	Proteína hipotética
	37,18	AMQ10702.1	Proteína contendo dedo de zinco
Golden marseillevirus	50,46	YP_009310305.1	Peptidase cisteína tipo papaína
	51,18	YP_009310318.1	Proteína hipotética
	51,44	YP_009310135.1	Proteína hipotética
	33,33	YP_009310173.1	Proteína hipotética
	34,23	YP_009310320.1	Proteína hipotética
	34,85	YP_009310191.1	Proteína hipotética
Lausannevirus	53,33	YP_004347316.1	Principal proteína de capsídeo
	53,64	YP_004347384.1	Proteína hipotética
	55,28	YP_004347348.1	Proteína de fusão histona H2B/H2A
	32,43	YP_004347287.1	Proteína hipotética
	34,23	YP_004347281.1	Proteína hipotética
	34,68	YP_004347324.1	Proteína hipotética
Tunisvirus	53,49	YP_009507112.1	Proteína hipotética
	54,01	YP_009506935.1	Proteína hipotética
	57,36	YP_009506948.1	Proteína hipotética
	31,15	YP_009507225.1	Proteína hipotética
	34,23	YP_009506995.1	Proteína hipotética
	34,46	YP_009506989.1	Proteína hipotética

As CDS *outliers* de maior e menor valor dentre os genomas de asfarvírus foram identificadas nos genomas de african swine fever vírus e pacmanvírus S19, respectivamente, com 62,96% e 19,66%. A maioria das CDS *outliers* no grupo foi identificada como proteína hipotética.

**Tabela 6. Valores GC% máximos e mínimos observados em *Asfarviridae*.
Asfarviridae - CDS máximas e mínimas**

Vírus	GC%	ID da proteína predita	Descrição
African swine fever virus	47,39	NP_042821.1	pE146L
	49,43	NP_042811.1	Precursor de pH233R
	62,96	NP_042844.1	pDP71L
	21,43	NP_042751.1	Proteína tipo lectina
	25,83	NP_042797.1	pD79L
	26,37	NP_042828.1	pE66L
Pacmanvirus S19	50,77	QYB17652.1	Proteína hipotética
	51,48	QYB17369.1	Proteína hipotética
	61,4	QYB17685.1	Proteína hipotética
	19,66	QYB17400.1	Proteína hipotética
	24,36	QYB17662.1	Proteína hipotética
	24,88	QYB17371.1	Proteína hipotética
Faustovirus e12	48,56	AIB52136.1	Proteína hipotética
	52,2	AIB52099.1	Proteína hipotética
	54,55	AIB52120.1	Proteína hipotética
	23,61	AIB52018.1	Proteína hipotética
	26,03	AIB52066.1	Proteína hipotética
	26,67	AIB52161.1	Proteína hipotética
Kaumobebavirus	52,65	QQV29307.1	Proteína hipotética
	53,12	QQV29569.1	Proteína hipotética
	53,33	QQV29750.1	Proteína hipotética
	30,6	QQV29493.1	Proteína hipotética
	33,12	QQV29492.1	Proteína hipotética
	33,25	QQV29754.1	Proteína hipotética

As regiões codificadoras de maior e menor valor GC% observadas em membros da família *Poxviridae* (**tabelas 7, 8 e 9**) são de, respectivamente, orf vírus com 83,91% e Amsacta moorei entomopoxvírus com 8,13%. Ambas as CDS anotadas são descritas como proteínas hipotéticas.

A maioria das CDS *outliers* em membros da família *Poxviridae* apresentaram descrições para proteínas hipotéticas e nenhuma tendência foi observada no grupo. Apesar disto, foi observada uma variedade de descrições para as CDS *outliers* do grupo, dentre os quais estão proteína da família EFC (complexo fusão e penetração), proteínas de membrana, antagonistas do receptor de IL-1 (apenas a porção N-terminal), fatores de virulência putativos, proteínas de ligação a TNF (“*tumor necrosis fator*”), etc.

Tabela 7. Tabela A de Poxviridae. Valores GC% máximos e mínimos observados em Poxviridae.

***Poxviridae* - CDS máximas e mínimas**

Vírus	GC%	ID da proteína predita	Descrição
Fowlpox virus	48,13	NP_039222.1	Proteína hipotética
	48,78	NP_038968.1	Proteína da família EFC (complexo fusão e entrada)
	48,78	NP_039219.1	Proteína da família EFC (complexo fusão e entrada)
	21,27	NP_039188.1	Homólogo a B20R de Vaccinia
	22,39	NP_039173.1	Proteína hipotética
	22,83	NP_039129.1	Proteína REDOX
Sheeppox virus	33,96	NP_659577.1	Proteína hipotética
	33,96	NP_659724.1	Proteína hipotética
	34,03	NP_659677.1	Proteína de membrana IMV
	16,67	NP_659583.1	Proteína hipotética
	18,1	NP_659700.1	Proteína semelhante a CD47
	18,15	NP_659589.1	Fator de crescimento semelhante ao EGF
Yokapox virus	33,33	YP_004821470.1	Proteína essencial de membrana IMV
	33,33	YP_004821539.1	Antagonista do receptor de IL-1
	38,31	YP_004821519.1	Proteína hipotética
	17,08	YP_004821506.1	Proteína hipotética
	18,14	YP_004821533.1	Proteína truncada grande tipo anquirina
	19,25	YP_004821363.1	Antagonista do receptor de IL-1 (apenas N-terminal)
Mule deerpox virus	37,42	ABI98989.1	Proteína hipotética
	37,42	ABI99155.1	Proteína hipotética
	38,69	ABI99116.1	Glicoproteína EEV
	19,48	ABI99114.1	Proteína hipotética
	20,21	ABI99023.1	Proteína hipotética
	20,48	ABI99074.1	Proteína de virion
Nile crocodilepox virus	74,6	YP_784228.1	Proteína IEV
	75,02	YP_784239.1	Proteína hipotética
	75,98	YP_784335.1	Resolvase de junção holliday
	43,25	YP_784338.1	Proteína semelhante à IMV A P4c de inclusão
	44,72	YP_784223.1	Proteína hipotética
	45,83	YP_784360.1	Proteína hipotética
Myxoma virus	54,17	NP_051878.1	m3.1L/R
	55,15	NP_051713.1	m2
	55,15	NP_051879.1	M-T2
	32,1	NP_051818.1	Proteína semelhante ao receptor de membrana IMV
	33,33	NP_051751.1	m37L
	34,33	NP_051732.1	Proteína hipotética
Eastern kangaroopox virus	61,59	ATI21112.1	Proteína hipotética
	61,74	ATI21221.1	Proteína tipo dedo de Zinco
	62,41	ATI21234.1	Proteína hipotética
	32,05	ATI21114.1	Proteína hipotética
	42,94	ATI21113.1	Proteína hipotética
	45,52	ATI21249.1	Proteína hipotética

Tabela 8. Tabela B de *Poxviridae*. Valores GC% máximos e mínimos observados em *Poxviridae*.

<i>Poxviridae</i> - CDS máximas e mínimas			
Vírus	GC%	ID da proteína predita	Descrição
Molluscum contagiosum virus	72,74	AZT86268.1	MC092R
	73,71	AZT86348.1	MC105L
	75,98	AZT86350.1	MC116R
	48,2	AZT86346.1	MC010R
	52,85	AZT86328.1	MC150R
	53,65	AZT86330.1	MC148R
Sea otterpox virus	40,92	YP_009480573.1	Proteína semelhante ao receptor de TNF
	41,05	YP_009480540.1	Proteína hipotética
	41,05	YP_009480669.1	Proteína hipotética
	22,84	YP_009480642.1	Proteína de membrana IMV não essencial
	22,92	YP_009480589.1	Componente de complexo de entrada/fusão
	23,02	YP_009480580.1	Componente de complexo de entrada/fusão
Vaccinia virus	44,63	YP_233100.1	Proteína de ligação a quimiocinas
	45,66	YP_232885.1	Proteína hipotética
	45,66	YP_233098.1	Proteína hipotética
	26,9	YP_233083.1	Proteína hipotética
	26,94	YP_233075.1	Proteína hipotética
	27,25	YP_232936.1	Proteína hipotética
Cotia virus	34,98	AIT70641.1	Proteína de ligação complementar
	35,98	AIT70781.1	Inibidor de NFkB
	42,11	AIT70637.1	Proteína de repetição de anquirina
	13,41	AIT70692.1	Proteína de formação crescente de montagem de vírus
	16,42	AIT70761.1	Proteína semelhante a CD47
	16,44	AIT70662.1	Proteína hipotética
Orf virus	77,78	NP_957781.1	Proteína hipotética
	77,78	NP_957910.1	Proteína hipotética
	83,91	NP_957782.1	Proteína hipotética
	38,3	NP_957880.1	Proteína de inclusão
	40,22	NP_957909.1	Proteína semelhante ao fator de crescimento endotelial vascular
	41,41	NP_957886.1	Glicoproteína EEV
Pteropox virus	48,72	YP_009268717.1	Proteína hipotética
	63,86	YP_009268716.1	Proteína hipotética
	63,86	YP_009268858.1	Proteína hipotética
	24,11	YP_009268762.1	Componente de complexo de entrada/fusão
	25,11	YP_009268748.1	Proteína hipotética
	25,32	YP_009268742.1	Serina recombinase não funcional conservada
Salmon gillpox virus	50,44	YP_009162575.1	Proteína hipotética
	53,33	YP_009162566.1	Proteína hipotética
	55,46	YP_009162382.1	Proteína hipotética de baixa complexidade
	26,13	YP_009162381.1	Proteína hipotética
	26,36	YP_009162544.1	Proteína hipotética
	27,23	YP_009162457.1	Proteína de baixa complexidade
Squirrelpox virus	75,66	YP_008658525.1	Proteína central (core)
	76,73	YP_008658499.1	Proteína hipotética conservada de poxvirus
	77,1	YP_008658496.1	Proteína hipotética conservada de poxvirus
	50	YP_008658528.1	Proteína de membrana IMV / fator de virulência putativo
	58,43	YP_008658556.1	Semelhante à proteína de ligação à quimiocina CC putativa
	58,67	YP_008658472.1	Proteína glutaredoxina-2

Tabela 9. Tabela C de *Poxviridae*. Valores GC% máximos e mínimos observados em *Poxviridae*.

<i>Poxviridae</i> - CDS máximas e mínimas			
Vírus	GC%	ID da proteína predita	Descrição
Swinepox virus	34,88	NP_570310.1	Proteína hipotética
	34,9	NP_570163.1	Proteína de ligação ao TNF do tipo MHC
	34,9	NP_570308.1	Proteína de ligação ao TNF do tipo MHC
	20,35	NP_570218.1	Proteína hipotética
	21,9	NP_570238.1	Proteína putativa de virion SPV078
	22,36	NP_570167.1	Proteína hipotética
Eptesipox virus	32,71	YP_009408134.1	Proteína hipotética
	33,33	YP_009408067.1	Proteína tipo dedo de Zinco, morfogênese tardia
	34,19	YP_009408123.1	Precursor da endotelina
	12,7	YP_009408127.1	Proteína hipotética
	14,22	YP_009408096.1	Proteína hipotética
	14,71	YP_009408109.1	Antagonista do receptor de IL-1
Yaba monkey tumor virus	39	NP_938267.1	Proteína hipotética
	45,77	NP_938402.1	Proteína hipotética
	45,87	NP_938263.1	Proteína hipotética
	20	NP_938376.1	Proteína hipotética
	22,22	NP_938361.1	106L
	23,47	NP_938384.1	Proteína semelhante a CD47
Anomala cuprea entomopoxvirus	34,55	YP_009001662.1	Proteína hipotética
	34,64	YP_009001497.1	Proteína hipotética
	36,78	YP_009001496.1	Proteína hipotética
	12,28	YP_009001519.1	Proteína hipotética
	12,99	YP_009001654.1	Proteína hipotética
	13,18	YP_009001511.1	Proteína hipotética
Amsacta moorei entomopoxvirus	34,85	NP_064870.1	Proteína hipotética
	36,02	NP_064972.1	Proteína hipotética
	38,96	NP_064973.1	Proteína hipotética
	8,13	NP_065034.1	Proteína hipotética
	10,08	NP_064899.1	Proteína hipotética
	10,53	NP_065011.1	Semelhante à proteinase clp dependente de ATP
Melanoplus sanguinipes entomopoxvirus	33,58	NP_048128.1	Proteína hipotética
	34,83	NP_048145.1	Proteína hipotética
	35,53	NP_048146.1	Proteína hipotética
	11,59	NP_048290.1	Proteína hipotética
	11,67	NP_048106.1	Proteína hipotética
	11,79	NP_048282.1	Proteína hipotética
Diachasmimorpha longicaudata entomopoxvirus	39,47	AKS26368.1	Proteína hipotética
	40,96	AKS26452.1	Proteína hipotética
	41,57	AKS26304.1	Proteína hipotética
	23,81	AKS26411.1	Proteína hipotética
	24,71	AKS26292.1	Proteína hipotética
	24,71	AKS26484.1	Proteína hipotética

As regiões codificadoras de maior e menor valor GC% observadas em membros do táxon “Pithoviridae” são de, respectivamente, pithovírus sibericum com 57,83% e orpheovírus com 9,18%. A CDS de maior GC% codifica uma proteína de repetição de tripla hélice de colágeno, enquanto a de menor valor GC% é descrita como proteína hipotética.

Não foram observados padrões em “Pithoviridae”. Entretanto, observamos proteínas contendo domínio transmembrana total e parcial, proteínas do tipo colágeno, proteína contendo domínio de colar de cauda de fago, dentre outras.

Tabela 10. Valores GC% máximos e mínimos observados em “Pithoviridae”.

Pithoviridae - CDS máximas e mínimas

Vírus	GC%	ID da proteína predita	Descrição
Pithovirus sibericum	55	YP_009001350.1	Proteína hipotética
	55,35	YP_009000908.1	Proteína hipotética
	57,83	YP_009001336.1	Proteína de repetição de tripla hélice de colágeno
	27,12	YP_009001041.1	Proteína hipotética
	27,73	YP_009001033.1	Proteína hipotética
	30,12	YP_009000986.1	Proteína hipotética
Cedratvirus A11	53,12	YP_009329357.1	Proteína contendo domínio BclB, Exosporium C-terminal, parcial
	53,6	YP_009329475.1	Proteína hipotética
	54,96	YP_009328963.1	Proteína hipotética
	34,57	YP_009329318.1	Proteína hipotética
	36,18	YP_009329147.1	Proteína hipotética
	37,48	YP_009329391.1	Proteína com proteína de membrana conservada e de função não caracterizada
Orpheovirus	43,58	YP_009449212.1	Proteína contendo o domínio do fator de iniciação da transcrição específica de RNA polimerase I
	44,13	YP_009448288.1	Proteína tipo colágeno
	46,92	YP_009449159.1	Proteína contendo domínio de colar de cauda de fago
	9,18	YP_009448976.1	Proteína hipotética
	11,11	YP_009448795.1	Proteína com domínio transmembrana
	12,17	YP_009449313.1	Proteína com domínio transmembrana, parcial

6. DISCUSSÃO

7.1 Extração de informações biológicas por análises *in silico* e a obtenção do percentual de GC de regiões específicas

Atualmente, com um acesso cada vez mais difuso e facilitado a um enorme contingente de informações biológicas, surgem novas oportunidades de olhar para um mesmo conteúdo e extrair deste algo novo. O estudo de regiões ricas em GC%, nos mais diversos organismos e grupos taxonômicos, pode ser considerado vasto e diverso por si próprio. Neste trabalho, pela primeira vez até, dentro do nosso conhecimento, o perfil da variação de conteúdo GC% de membros do filo *Nucleocytoviricota* foi mapeado, gene a gene, permitindo-nos olhar para estes vírus tão estudados a partir de um ponto de vista diferente.

Por meio de abordagens como esta, é possível observar com mais clareza características específicas das famílias representadas neste grupo, e a partir destas descrever mais sobre a biologia destes microrganismos. Em determinados grupos dentro *Nucleocytoviricota*, como é o caso do táxon *Mimiviridae* – um grupo ainda recente e menos estudado, proporcionalmente a outros grupos próximos – trabalhos que possam descrever tendências evolutivas ou características destes imensos genomas tornam-se ainda mais importantes.

Uma outra qualidade das análises *in silico*, que utilizam dados biológicos depositados em bancos de dados públicos, é a conveniência de que qualquer pesquisador com acesso a estes conteúdos e a ferramentas informáticas simples pode explorar este mundo de dados, de maneira simples ou não, e colaborar com a infinita construção do conhecimento biológico.

7.2 Variação global de GC% em *Nucleocytoviricota*

Em membros do filo *Nucleocytoviricota*, observamos uma amplitude de valores absolutos de GC% de sequências codificantes de 8,13%, referente à CDS NP_065034.1 de *Amsacta moorei* entomopoxvirus, e de 83,91%, referente à CDS NP_957782.1 de orf vírus, ambas anotadas como proteínas hipotéticas. Interessantemente, os genomas aos quais os valores mínimos e máximos de percentual GC pertencem à família *Poxviridae*. Levando em consideração o conjunto de dados adotados nestas análises, *Poxviridae* ilustra a variação de GC% em *Nucleocytoviricota*, mas enfatiza, também, a diversidade interna à família viral, contemplada por ambos os valores extremos observados.

Os valores médios de percentual GC relativo às regiões codificantes, por sua vez, vão de 19% a 67%, referentes aos respectivos genomas de *Melanoplus sanguinipes* entomopoxvírus e *Squirrelpox* vírus, o que reforça a diversidade dentro do grupo mais uma vez.

A média de GC% total dos genomas, quando comparada com a média de GC% das regiões codificantes, não apresentou grandes oscilações, o que indica que não há, a princípio, nenhuma tendência evolutiva que torne significativamente distinta a composição das sequências intergênicas e das sequências codificantes em relação ao conteúdo GC%. Grupos para os quais foi possível observar alguma variação entre os valores médios genômicos e os valores médios de regiões codificantes foram os membros dos *taxa Iridoviridae* e “*Pithoviridae*”, considerando ainda que este último inclui o genoma de orpheovírus, um *pithoviridae*-like. No caso destes dois grupos, especificamente, é importante observar que as diferenças nas médias não foram avaliadas em relação à sua significância estatística e, além disto, os grupos possuem um número menor de genomas incluídos neste trabalho.

Como dito anteriormente, *Nucleocytoviricota* é um grupo taxonômico muito variado e o banco de dados montado e utilizado neste trabalho, apesar de contemplar parte desta diversidade, deixa de fora diversos outros integrantes virais que poderiam agregar robustez às análises realizadas. Ainda, para que os valores médios de membros de *Nucleocytoviricota* aproximassem-se mais da realidade, deveria ser também

considerado o viés de estudo – que, neste caso, supra representa *Poxviridae*, a família de vírus nucleocitoplasmáticos com grande relevância médica humana e veterinária.

7.3 Variação de GC% dentro das famílias virais

A variação de percentual GC dentre os membros de *Phycodnaviridae* é elevada, principalmente considerando que, neste trabalho, o grupo está subrepresentado. Inicialmente, observando os mapas dos taxa *Phycodnaviridae* e de *Nucleocytoviricota* (**figuras 16 e 23**), os genomas de Ectocarpus siliculosus vírus 1 e Heterosigma akashiwo vírus 1 chamam atenção por terem a maior e a menor média de GC% do grupo, respectivamente.

Outra característica observável que chama atenção está no genoma de Emiliana huxleyi vírus 86, sendo esta um *cluster* localizado em torno da posição relativa de CDS 300 (**figura 6**), cujos valores estão notavelmente acima da média do genoma, tornando esta região um potencial alvo para estudos futuros. Tal elevada variação de GC% dentre os ficodnavírus analisados pode ser parcialmente explicada pelo fato de que este táxon é sabidamente polifilético, contendo representantes que deverão ser alocadas em outros taxa virais, como *Mimiviridae* e “*Pandoraviridae*”.

Observando os mapas de perfil de variação de GC% em conjunto, ou seja, o mapa de *Nucleocytoviricota* (**figura 23**), alguns aspectos logo chamam a atenção. A família *Mimiviridae*, com 13 genomas representantes, destaca-se pela coloração azul significativamente maior em relação às demais famílias representadas, caracterizando a como uma família viral de genomas cujo percentual GC é tradicionalmente baixo, o que corrobora com o que se observa descrito na literatura (“*Mimiviridae*,” 2012).

Tratando-se do táxon *Mimiviridae*, ainda, pode-se observar que os genomas mantêm uma certa equivalência para com a média de GC% do grupo – o oposto do que se observa para *Poxviridae*, a ser discutido posteriormente – com exceção dos genomas Cafeteria roenbergensis vírus e Tetraselmis viridis vírus, respectivamente abaixo e acima da média do grupo, em um aspecto geral.

Os genomas particularmente grandes representados neste grupo também chamam atenção pela presença de regiões com valores GC% com notável oscilação, como é o caso de *Chrysochromulina ericina* vírus, que possui um conjunto de CDS de posição relativa ao final do genoma, cujo GC% é significativamente maior em relação à média do próprio vírus e em relação à média GC% de *Mimiviridae* (**figura 7**).

Clusters de CDS com valor percentual discrepante também são observados para outros genomas de *Mimiviridae*, como é o caso de *cotonvirus japonicus*, com variações acima da média dispostas entre as marcas relativas das posições das CDS de 400 e 600, e na porção final do genoma (**Figura 8**).

Ainda, é possível observar que, de modo geral, os genomas representados possuem regiões codificantes com valores GC% bastante próximos da média, com exceções, como *Phaeocystis globosa* vírus, por exemplo, cuja variação interna de valores de GC% é maior (**figura 7**).

Os genomas do táxon *Iridoviridae* representados neste trabalho apresentam uma considerável consistência em relação à variação do conteúdo GC% dentro de cada genoma (**figura 9**). No entanto, quando comparados entre si, evidencia-se que dois dos genomas estudados são bastante diferentes, sendo estes *invertebrate iridescente* vírus 3 e *frog* vírus 3. Ambos os genomas apresentam uma média GC% elevada.

Em *Iridoviridae*, foram observadas diferenças entre as médias de GC% genômico e GC% médio das CDS, o que não foi observado para a maioria das famílias estudadas neste trabalho. Embora não tenham sido realizados testes estatísticos de significância, podemos chamar atenção para os genomas de *invertebrate iridescent* vírus 6 (28% e 35%) e *decapod iridescent* vírus 1 (34,6% e 29,8%).

O táxon *Marseilleviridae*, diferente dos grupos anteriormente citados, apresentou pouca variação de valores GC%, tanto dentro de cada genoma individual, quanto como grupo (**figuras 10 e 19**). O perfil de distribuição GC% em *Marseilleviridae* é, portanto, bem conservado no grupo, como já tinha sido observado por Blanca e colaboradores, embora o trabalho mencionado tivesse a expectativa de observar resultados distintos. Isto porque os genomas dos *marseillevírus* são circulares e apresentam-se de forma

circular e assimétricos – uma região tende a concentrar os denominados “*core genes*”, que são altamente conservados, enquanto a outra região apresenta uma grande proporção de parálogos sob grande influência de rearranjos genômicos (BLANCA *et al.*, 2020).

No táxon *Asfarviridae*, levando em consideração os três genomas do grupo conhecido como extended *Asfarviridae*, observamos genomas com grande amplitude de valores GC%, como pacmanvírus S19 e african swine fever vírus (ASFV), e genomas com baixa amplitude de valores GC%, como kaumoebavírus (**figuras 11 e 20**). Com médias individuais similares, o grupo também se apresenta consistente em relação à média do filo. No entanto, dois dos genomas estudados apresentam uma variação maior no GC% de regiões codificantes, sendo estes pacmanvírus S19 e kaumoebavírus, respectivamente com maior proporção de CDS abaixo e acima da média dos membros de *Nucleocyotoviricota* e de *Asfarviridae*.

A variação de GC% ao longo do genoma de ASFV apresenta, ainda, extremidades do genoma com regiões codificantes de valores abaixo da média, enquanto a porção interna apresenta uma tendência a valores de GC% acima da média.

Neste trabalho, as maiores variações de GC% foram observadas em *Poxviridae*. Como mencionado anteriormente, nesta família viral estão os extremos de valores absolutos e médias de conteúdo GC%, o que pode caracterizar os poxvírus como vírus que possuem um perfil de GC% muito variável.

Observando os genomas dos poxvírus em relação aos demais genomas de NCLDV's analisados neste trabalho, três grupos se destacam. O primeiro grupo, composto pelos genomas de Nile crocodilepox vírus, eastern kangaropox vírus, molluscum contagiosum vírus, orf vírus e squirrelpox vírus, apresenta genomas com notáveis elevados valores de GC%. O segundo grupo, composto por fowlpox vírus, sheeppox vírus, yokapox vírus, mule deerpox vírus, Cotia vírus, swinepox vírus e eptesipox vírus, ao contrário do primeiro, destaca-se por valores de GC% muito abaixo da média. O terceiro grupo, composto pelos vírus da subfamília *Entomopoxvirinae*, também se destaca por valores GC% baixos, porém ainda mais abaixo dos demais poxvírus citados.

7.4 Alvos para trabalhos futuros

Diferentes abordagens e métodos podem ser utilizados a fim de identificar novos alvos em sequências biológicas a serem estudados, especialmente porque as tecnologias recentes de sequenciamento e a disponibilização de dados biológicos configuram um cenário propício para a caracterização de genomas – incluindo os genomas virais. Utilizando do método descrito neste trabalho, observamos diferentes características individuais dos vírus do táxon *Nucleocytoviricota*, e também características das famílias pertencentes ao filo. Algumas destas características observadas podem nos levar a potenciais novos alvos para estudos futuros, como será sugerido a seguir.

Considerando vírus do táxon *Phycodnaviridae*, como mencionado em seções anteriores, o genoma de *Emiliana huxleyi* vírus 86 (EhV-86), com 407,34 Kb e GC% médio total de 40,2% (**tabela 1**), apresenta um bloco de regiões codificantes particularmente distinto do restante do genoma, entre as CDS de posição relativa 250 e 350 (**figura 6**). Esta região do genoma em questão pode trazer informações relevantes para a biologia do vírus, ou mesmo elucidar aspectos sobre a evolução deste e de seu hospedeiro – no caso, por exemplo, deste ser um *cluster* de sequências adquiridas via transferência gênica horizontal, um evento biológico muito observado entre vírus e seus hospedeiros, ou mesmo entre diferentes microrganismos em um mesmo hospedeiro. Este é um evento já previamente descrito para o EhV-86, segundo Adam Monier e colaboradores (2009), que demonstram a transferência horizontal da via metabólica completa da biossíntese de esfingolípídeos e ceramidas entre o vírus e seu hospedeiro, *Emiliana huxleyi*, uma alga unicelular.

De maneira similar, os genomas de *Chrysochromulina ericina* vírus, megavirus *chiliensis*, cotonvirus *japonucus* e Bodo saltans vírus também apresentam pequenos agrupamentos de CDS com valores discrepantes em relação à média individual de cada genoma e à média do grupo *Mimiviridae* (**figura 17**). Além disto, os genomas grandes e com significativa quantidade de genes que observamos nesta família viral, em conjunto dos fatores biológicos associados a seus hospedeiros – como é o caso das amebas,

conhecidas como verdadeiros cavalos de troia do mundo microbiano por abrigarem uma variedade de organismos em seu interior – tornam propício o evento da transferência gênica horizontal. Ainda, observamos que a busca pela descrição das CDS de maiores e menores valores em *Mimiviridae* retornou uma quantidade muito elevada de ORFans e proteínas hipotéticas.

Outras regiões que concentram CDS de interesse para estudo futuros estão representadas em *Poxviridae* em um perfil bastante característico. No atual trabalho, os genomas de poxvírus apresentam, em sua maioria, um padrão no qual as regiões codificantes das extremidades aparentam ser repetidas e invertidas – isto é, espelhadas. Nos genomas de fowlpox vírus, mule deerpox vírus, myxoma vírus, sea otterpox vírus, vaccínia vírus, Cotia vírus, pteropox vírus, orf vírus, swinepox vírus, epstesipox vírus e *Melanoplus sanguinipes* entomopoxvírus este padrão está ainda mais evidente (**figura 21**).

O padrão descrito acima e aqui observado para os poxvírus corrobora com aspectos da biologia de membros da família *Poxviridae*, ilustrando por meio do perfil de GC% gene a gene as denominas regiões terminais invertidas repetidas (ou “*inverted terminal repeats – ITR*”), que são regiões tradicionalmente consideradas alvos em estudos sobre poxvírus (WITTEK *et al.*, 1978).

7. CONCLUSÃO

O estudo de composição de sequências biológicas, como DNA, RNA e sequências proteicas, consiste de uma das abordagens recentes mais inovadoras para inferência de aspectos variados da biologia dos organismos de todos os domínios da vida, incluindo os vírus.

Neste trabalho, utilizando de ferramentas simples e de relativo fácil acesso, bem como acessando dados públicos, foi possível descrever brevemente o perfil global do percentual GC, guanina e citosina, de um dos mais diversos grupos taxonômicos atualmente conhecidos, *Nucleocytoviricota*. Não somente foram descritos os perfis de variação de conteúdo GC% de algumas famílias integrantes do filo – *Phycodnaviridae*, *Mimiviridae*, *Iridoviridae*, *Marseilleviridae*, *Poxviridae* e “*Pithoviridae*” – como também, pela primeira vez, foi descrito o perfil de GC% de todas as regiões codificantes conhecidas nos genomas estudados.

REFERÊNCIAS BIBLIOGRÁFICAS

ABERGEL, Chantal; CLAVERIE, Jean Michel. **Pithovirus sibericum: réveil d'un virus géant de plus de 30 000 ans**. *Medecine/Sciences*, v. 30, n. 3, p. 329–331, 2014.

ABRAHÃO, Jônatas *et al.* **Tailed giant Tupanvirus possesses the most complete translational apparatus of the known virosphere**. *Nature Communications*, v. 9, n. 1, 2018a.

ANDRADE, Ana Cláudia Dos S.P. *et al.* **Ubiquitous giants: A plethora of giant viruses found in Brazil and Antarctica**. *Virology Journal*, v. 15, n. 1, 24 jan. 2018.

BELLINI, William J; ROTA, Jennifer S; ROTA, Paul A. ***Virology of Measles Virus***. . [S.l: s.n.], [S.d.]. Disponível em: <<http://jid.oxfordjournals.org/>>.

BENSON, Dennis A *et al.* ***GenBank***. *Nucleic Acids Research*. [S.l: s.n.], 2000. Disponível em: <<http://www.ncbi.nlm.nih.gov/Entrez/Genome/>>.

BLANCA, Léo *et al.* **Comparative analysis of the circular and highly asymmetrical marseilleviridae genomes**. *Viruses*, v. 12, n. 11, 1 nov. 2020.

BOS, Lute. **100 years of virology : from vitalism via molecular biology to genetic engineering**. p. 82–87, 1888.

BOUVIER, Nicole M.; PALESE, Peter. **The biology of influenza viruses**. *Vaccine*, v. 26, n. SUPPL. 4, 12 set. 2008.

FILÉE, Jonathan. ***Route of NCLDV evolution: The genomic accordion***. *Current Opinion in Virology*. [S.l.]: Elsevier B.V. , 2013

FILÉE, Jonathan; CHANDLER, Michael. **Gene exchange and the origin of giant viruses**. *Intervirology*, v. 53, n. 5, p. 354–361, 2010.

FISHER, Christine R.; STREICKER, Daniel G.; SCHNELL, Matthias J. **The spread and evolution of rabies virus: Conquering new frontiers**. *Nature Reviews Microbiology*. [S.l.]: Nature Publishing Group. , 1 abr. 2018

FORTERRE, Patrick. **Giant viruses: Conflicts in revisiting the virus concept**. *Intervirology*. [S.l.]: S. Karger AG. , 2010

FRADA, Miguel *et al.* **The “Cheshire Cat” escape strategy of the coccolithophore *Emiliania huxleyi* in response to viral infection**. . [S.l: s.n.], 2008. Disponível em: <www.pnas.org/cgi/doi/10.1073/pnas.0807707105>.

GREUB, Gilbert; RAOULT, Didier. **Microorganisms Resistant to Free-Living *Amoebae***. *Clinical Microbiology Reviews*. [S.l: s.n.], , abr. 2004

IYER, Lakshminarayan M. *et al.* **Evolutionary genomics of nucleo-cytoplasmic large DNA viruses**. *Virus Research*, v. 117, n. 1, p. 156–184, 2006.

IYER, Lakshminarayan M.; ARAVIND, L.; KOONIN, Eugene V. **Common Origin of Four Diverse Families of Large Eukaryotic DNA Viruses**. *Journal of Virology*, v. 75, n. 23, p. 11720–11734, 2001.

JAUHAL, April A.; NEWCOMB, Richard D. **Assessing genome assembly quality prior to downstream analysis: N50 versus BUSCO**. *Molecular Ecology Resources*, v. 21, n. 5, p. 1416–1421, 1 jul. 2021.

KHRUSTALEV, Vladislav Victorovich; BARKOVSKY, Eugene Victorovich. **Mutational pressure is a cause of inter- and intragenomic differences in GC-content of simplex and varicello viruses**. *Computational Biology and Chemistry*, v. 33, n. 4, p. 295–302, ago. 2009.

KIMURA, Makoto *et al.* **Ecology of viruses in soils: Past, present and future perspectives**. *Soil Science and Plant Nutrition*. [S.l: s.n.], , fev. 2008

KLUG, A. **The tobacco mosaic virus particle: Structure and assembly**. *Philosophical Transactions of the Royal Society B: Biological Sciences*, v. 354, n. 1383, p. 531–535, 1999.

LA SCOLA, Bernard *et al.* **The virophage as a unique parasite of the giant mimivirus**. *Nature*, v. 455, n. 7209, p. 100–104, 4 set. 2008.

LECOQ, Hervé. Découverte du premier virus , **le virus de la mosaïque du tabac** : 1892 ou 1898 ? v. 324, p. 929–933, 2001.

LEGENDRE, Matthieu *et al.* **In-depth study of Mollivirus sibericum, a new 30,000-yold giant virus infecting Acanthamoeba**. *Proceedings of the National Academy of Sciences of the United States of America*, v. 112, n. 38, p. E5327–E5335, 22 set. 2015.

LEGENDRE, Matthieu *et al.* **Thirty-thousand-year-old distant relative of giant icosahedral DNA viruses with a pandoravirus morphology**. *Proceedings of the National Academy of Sciences of the United States of America*, v. 111, n. 11, p. 4274–4279, 2014.

LI, Yu *et al.* **GC content-based pan-pox universal PCR assays for poxvirus detection**. *Journal of Clinical Microbiology*, v. 48, n. 1, p. 268–276, jan. 2010.

LOBO, Francisco P. *et al.* **Virus-host coevolution: Common patterns of nucleotide motif usage in Flaviviridae and their hosts**. *PLoS ONE*, v. 4, n. 7, 20 jul. 2009.

LUSTIG, Alice; LEVINE, Arnold J. MINIREVIEW **One Hundred Years of Virology**. v. 66, n. 8, p. 4629–4631, 1992.

LWOFF, BY A. **The Concept of Virus The Third Marjory Stephenson Memorial Lecture**. *Journal of General Microbiology was issued on*. [S.l: s.n.], 1957.

MANNI, Mosè *et al.* **BUSCO Update: Novel and Streamlined Workflows along with Broader and Deeper Phylogenetic Coverage for Scoring of Eukaryotic, Prokaryotic, and Viral Genomes**. *Molecular Biology and Evolution*, v. 38, n. 10, p. 4647–4654, 1 out. 2021.

Mimiviridae. *Virus Taxonomy*. [S.l.]: Elsevier, 2012. p. 223–228.

MONATH, Thomas P. **Yellow fever: an update.** *THE LANCET Infectious Diseases*. [S.l: s.n.], 2001.

OUYANG, Qingjian *et al.* **High GC content of simple sequence repeats in Herpes simplex virus type 1 genome.** *Gene*, v. 499, n. 1, p. 37–40, 10 maio 2012.

PAREEK, Chandra Shekhar; SMOCZYNSKI, Rafal; TRETYN, Andrzej. **Sequencing technologies and genome sequencing.** *Journal of Applied Genetics*. [S.l: s.n.], , nov. 2011

PARK, Sang Tae; KIM, Jayoung. **Trends in next-generation sequencing and a new era for whole genome sequencing.** *International Neuourology Journal*. [S.l.]: Korean Continenence Society. , 2016

PURCELL, Robert H. **Hepatitis viruses: Changing patterns of human disease.** *Proc. Nati. Acad. Sci. USA*. [S.l: s.n.], 1994. Disponível em: <<https://www.pnas.org>>.

RAOULT, Didier; LA SCOLA, Bernard; BIRTLES, Richard. **The discovery and characterization of mimivirus, the largest known virus and putative pneumonia agent.** *Clinical Infectious Diseases*. [S.l: s.n.], , 1 jul. 2007

ROBINSON, Christopher M. *et al.* **Molecular evolution of human adenoviruses.** *Scientific Reports*, v. 3, 2013.

SCOLA, Bernard La *et al.* **A Giant Virus in Amoebae.** . [S.l: s.n.], [S.d.]. Disponível em: <www.ebi.ac.uk/clustalw/>.

SÉMON, Marie; MOUCHIROUD, Dominique; DURET, Laurent. **Relationship between gene expression and GC-content in mammals: Statistical significance and biological relevance.** *Human Molecular Genetics*, v. 14, n. 3, p. 421–427, 1 fev. 2005.

SUZAN-MONTI, M.; LA SCOLA, B.; RAOULT, D. **Genomic and evolutionary aspects of Mimivirus.** *Virus Research*. [S.l: s.n.], , abr. 2006

WEINBAUER, Markus G. **Ecology of prokaryotic viruses.** *FEMS Microbiology Reviews*. [S.l: s.n.], , maio 2004

WILHELM, Steven W; SUTTLE, Curtis A. **Viruses and Nutrient Cycles i n the Sea**
Viruses play critical roles in the structure and function of aquatic food webs
Distribution of bacteria and viruses in the sea. . [S.l: s.n.], [S.d.]. Disponível em: <<https://academic.oup.com/bioscience/article/49/10/781/222807>>.

WITTEK, Riccardo *et al.* **Inverted Terminal Repeats in Rabbit Poxvirus and Vaccinia Virus DNA.** *JOURNAL OF VIROLOGY*. [S.l: s.n.], 1978. Disponível em: <<https://journals.asm.org/journal/jvi>>.

WU, Hao *et al.* **On the molecular mechanism of GC content variation among eubacterial genomes.** *Biology Direct*, v. 7, 10 jan. 2012.

YOSHIKAWA, Genki *et al.* **Medusavirus, a Novel Large DNA Virus Discovered from Hot Spring Water.** *Journal of Virology*, v. 93, n. 8, 15 abr. 2019.

YUTIN, Natalya *et al.* **Eukaryotic large nucleo-cytoplasmic DNA viruses: Clusters of orthologous genes and reconstruction of viral genome evolution.** *Virology Journal*, v. 6, p. 1–13, 2009.

ZHANG, Qi Ya; GUI, Jian Fang. Diversity, evolutionary contribution and ecological roles of aquatic viruses. *Science China Life Sciences*. [S.l.]: Science in China Press. , 1 dez. 2018

PRODUÇÕES CIENTÍFICAS

I. Artigos aceitos para publicação

KALAPOTHAKIS, YAN ; MIRANDA, KELTON ; PEREIRA, ADRIANA HELOÍSA ; **WITT, AMANDA S.A.** ; MARANI, CAMILA ; MARTINS, ANA PAULA ; LEAL, HORTÊNSIA GOMES ; CAMPOS-JÚNIOR, EDIMAR ; PIMENTA, ADRIANO M.C. ; BORGES, ADOLFO ; CHÁVEZ-OLÓRTEGUI, CARLOS ; KALAPOTHAKIS, EVANGUEDES . **Novel components of *Tityus serrulatus* venom: A transcriptomic approach.** TOXICON, v. 189, p. 91-104, 2021. (Artigo publicado)

Amanda Stéphanie Arantes Witt, Ana Paula Correia Crispim, Jônatas Santos Abrahão, Viviane de Souza Alves. **Sobre as origens, história e consequências do negacionismo científico.** PROJETO E-BOOK SOBRE NEGACIONISMO DA CIÊNCIA. Iniciativa: PUC Minas, em parceria com a Rede Mineira de Comunicação Científica (RMCC), 2021 (Artigo aceito para publicação).

BORATTO, PAULO VICTOR M.; SERAFIM, MATEUS SÁ M.; **WITT, AMANDA STÉPHANIE A.**; CRISPIM, ANA PAULA C.; AZEVEDO, BRUNA LUIZA DE; SOUZA, GABRIEL AUGUSTO P. DE; AQUINO, ISABELLA LUIZA M. DE; MACHADO, TALITA B.; QUEIROZ, VICTÓRIA F.; RODRIGUES, RODRIGO A. L.; BERGIER, IVAN; CORTINES, JULIANA REIS; FARIAS, SAVIO TORRES DE; SANTOS, RAÍSSA NUNES DOS; CAMPOS, FABRÍCIO SOUZA; FRANCO, ANA CLÁUDIA; ABRAHÃO, JÔNATAS S. A Brief History of Giant Viruses. Studies in Brazilian Biomes. Viruses-Basel **JCR**, v. 14, p. 191, 2022. (Artigo publicado)

II. Resumo publicado em evento científico

WITT, A. S. A.; ABRAHÃO, JÔNATAS S. VARIAÇÃO DE PERFIL DE CONTEÚDO GC% DE REGIÕES CODIFICANTES EM NUCLEOCYTOVIRICOTA. In: VIII Simpósio de Microbiologia da UFMG - CONECTA SIM 2021, 2021, Belo Horizonte. VIII Simpósio de Microbiologia da UFMG - CONECTA SIM 2021, 2021.