MINISTÉRIO DA SAÚDE
FUNDAÇÃO OSWALDO CRUZ
INSTITUTO OSWALDO CRUZ

Mestrado em Programa de Pós-Graduação *stricto sensu*
em Biologia Computacional e Sistemas

DESENVOLVIMENTO DE MODELOS ESTATÍSTICOS PARA ENSAIOS FENOTÍPICOS AUTOMATIZADOS DE *SCHISTOSOMA MANSONI* ADULTO

FÁBIO JORGE DE VASCONCELLOS JÚNIOR

Rio de Janeiro
Março de 2019

Ministério da Saúde

FIOCRUZ
Fundação Oswaldo Cruz

INSTITUTO OSWALDO CRUZ
Programa de Pós-Graduação
em Biologia Computacional e Sistemas

FÁBIO JORGE DE VASCONCELLOS JÚNIOR

Desenvolvimento de modelos estatísticos para ensaios fenotípicos automatizados de *Schistosoma mansoni* adulto

Dissertação apresentada ao Instituto Oswaldo Cruz como parte dos requisitos para obtenção do título de Mestre em Ciências (Biologia Computacional e Sistemas)

Orientador (es): Prof. Dr. Floriano Paes Silva Jr.
Prof. Dr. Leonardo Soares Bastos

RIO DE JANEIRO
Março de 2019

Jorge de Vasconcellos Júnior, Fábio.

Desenvolvimento de Modelos Estatísticos para Ensaios Fenotípicos Automatizados de Schistosoma *mansoni* adulto / Fábio Jorge de Vasconcellos Júnior. - Rio de Janeiro, 2019.
161 f.; il.

Dissertação (Mestrado) – Instituto Oswaldo Cruz, Pós-Graduação em Biologia Computacional e Sistemas, 2019.

Orientador: Floriano Paes Silva Júnior.
Co-orientador: Leonardo Soares Bastos.

Bibliografia: f. 111-122

1. esquistossomose. 2. ensaios fenotípicos. 3. modelagem estatística. I. Título.

Elaborada pelo Sistema de Geração Automática de Ficha Catalográfica da Biblioteca de Manguinhos/ICICT com os dados fornecidos pelo(a) autor(a).

Ministério da Saúde

FIOCRUZ
Fundação Oswaldo Cruz

INSTITUTO OSWALDO CRUZ
Programa de Pós-Graduação
em Biologia Computacional e Sistemas

AUTOR: FÁBIO JORGE DE VASCONCELLOS JÚNIOR

DESENVOLVIMENTO DE MODELOS ESTATÍSTICOS PARA ENSAIOS FENOTÍPICOS AUTOMATIZADOS DE *SCHISTOSOMA MANSONI* ADULTO

ORIENTADOR(ES): **Prof. Dr. Floriano Paes Silva Jr.**
Prof. Dr. Leonardo Soares Bastos

Aprovada em: 29/03/2019

EXAMINADORES:

Prof. Dr. Fabrício Alves Barbosa da Silva - Presidente (PROCC- Fiocruz)
Prof. Dr. Gustavo da Silva Ferreira (ENCE - IBGE)
Prof. Dr. Marcelo Santos Castilho (Faculdade de Farmácia - UFBA)
Prof. Dr. Eduardo Caio Torres dos Santos (IOC - Fiocruz)
Prof. Dr. Marcelo Ferreira da Costa Gomes (PROCC - Fiocruz)

Rio de Janeiro, 29 de Março de 2019

Ministério da Saúde

Fundação Oswaldo Cruz
Instituto Oswaldo Cruz

Ata da defesa de dissertação de mestrado em Biologia Computacional e Sistemas de **Fabio Jorge de Vasconcellos Júnior**, sob orientação do Dr. Floriano Paes Silva Júnior e Dr. Leonardo Soares Bastos. Ao vigésimo nono dia do mês de março de dois mil e dezenove, realizou-se às dez horas, na Sala 1 - Módulo de Expansão do Ensino, o exame da dissertação de mestrado intitulada: **"Desenvolvimento de modelos estatísticos para ensaios fenotípicos automatizados de *Schistosoma mansoni* adulto"**, no programa de Pós-graduação em Biologia Computacional e Sistemas do Instituto Oswaldo Cruz, como parte dos requisitos para obtenção do título de Mestre em Ciências - área de concentração: Biologia Molecular Estrutural, na linha de pesquisa: Abordagens Computacionais no desenvolvimento de Fármacos e Vacinas. A banca examinadora foi constituída pelos Professores: Dr. Fabricio Alves Barbosa da Silva - IOC/FIOCRUZ (Presidente), Dr. Gustavo da Silva Ferreira - IBGE/RJ, Dr. Marcelo Santos Castilho - UFBA/BA e como suplentes: Dr. Eduardo Caio Torres dos Santos – IOC/FIOCRUZ e Dr. Marcelo Ferreira da Costa Gomes – PROCC/FIOCRUZ. Após arguir o candidato e considerando que o mesmo demonstrou capacidade no trato do tema escolhido e sistematização da apresentação dos dados, a banca examinadora pronunciou-se pela APROVAÇÃO da defesa da dissertação de mestrado. De acordo com o regulamento do Curso de Pós-Graduação em Biologia Computacional e Sistemas do Instituto Oswaldo Cruz, a outorga do título de Mestre em Ciências está condicionada à emissão de documento comprobatório de conclusão do curso. Uma vez encerrado o exame, a Coordenadora do Programa Drª. Ana Carolina Ramos Guimarães, assinou a presente ata tomando ciência da decisão dos membros da banca examinadora. Rio de Janeiro, 29 de março de 2019.

Dr. Fabricio Alves Barbosa da Silva (Presidente da Banca):

Dr. Gustavo da Silva Ferreira (Membro da Banca):

Dr. Marcelo Santos Castilho (Membro da Banca):

Drª. Ana Carolina Ramos Guimarães (Coordenadora do Programa):

Av. Brasil, 4365 Manguinhos Rio de Janeiro RJ Brasil **CEP:** 21040-360
Contatos: (21) 2562-1201 / 2562-1299 **E-mail:** atendimentoseac@ioc.fiocruz.br **Site:** www.fiocruz.br/iocensino

AGRADECIMENTOS

À Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – CAPES pelo auxílio financeiro, pois este trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Código de Financiamento 001.

Ao programa de Pós-Graduação em Biologia Computacional e Sistemas (IOC - Fiocruz), incluindo todos os docentes e discentes com quem pude interagir, pela oportunidade de desenvolvimento pessoal e profissional a mim oferecida.

Ao meu orientador, Floriano Paes Silva Jr., e coorientador, Leonardo Bastos, por toda ajuda e acompanhamento a mim oferecidos; por toda paciência que demonstraram possuir; e pela educação e elegância com que me trataram todo esse tempo de mestrado, mesmo quando eu talvez tivesse merecido uns tapas na cara.

A todos do grupo LaBECFar, por me receber tão carinhosamente. Em especial, ao Rafael Dantas, por toda a ajuda que me ofereceu prontamente.

Não poderia deixar de agradecer também à Thaís, minha companheira para toda a vida, pelo nosso amor, que me sustenta e ampara todos os dias.

Ministério da Saúde

FIOCRUZ
Fundação Oswaldo Cruz

INSTITUTO OSWALDO CRUZ

DESENVOLVIMENTO DE MODELOS ESTATÍSTICOS PARA ENSAIOS FENOTÍPICOS AUTOMATIZADOS DE *SCHISTOSOMA MANSONI* ADULTO

RESUMO

DISSERTAÇÃO DE MESTRADO EM BIOLOGIA COMPUTACIONAL E SISTEMAS

Fábio Jorge de Vasconcellos Júnior

Dentre as doenças parasitárias, a esquistossomose é um grave mal que assola parte da população mais pobre. Por isso, a pesquisa por novas terapias farmacológicas contra o seu agente etiológico (os parasitas do gênero *Schistosoma*) é uma atividade de grande relevância social e econômica. Entretanto, está atualmente disponível, na prática, um único fármaco, que já enfrenta problemas contra cepas resistentes. A primeira etapa no processo de pesquisa e o desenvolvimento de fármacos consiste na identificação de compostos com atividade biológica (*hits*), cujo método comumente empregado é a triagem de uma biblioteca de compostos baseada no alvo ou baseada em fenótipo. Contra as parasitoses, a abordagem fenotípica é muito relevante, graças ao seu maior valor biológico (sistêmico), comparado a um ensaio bioquímico. Além disso, como sua moderna extensão, estão disponíveis métodos automatizados de triagem de alto conteúdo (HCS) em células ou em pequenos organismos. Uma das dificuldades desta abordagem está na insuficiência das ferramentas estatísticas mais difundidas na análise de dados de maior complexidade. Logo, faz-se necessário o desenvolvimento de modelos para uma análise estatística adequada aos dados complexos obtidos experimentalmente, a fim de estimar o efeito do composto no fenótipo do parasito. Desta forma, o objetivo do presente trabalho foi o desenvolvimento de modelos estatísticos para a análise do efeito de compostos candidatos antiesquistossomais, a partir de dados quantitativos de ensaios fenotípicos automatizados, obtidos em *Schistosoma mansoni* adulto com a tecnologia HCS na plataforma de Bioensaios e Triagem de Fármacos do IOC/FIOCRUZ. A partir da análise estatística exploratória dos dados, pôde-se propor e comparar modelos para a inferência do efeito de cada composto testado, a partir de dados de motilidade do parasita. Foram avaliados modelos lineares generalizados multinível, em busca de adequação à estrutura hierárquica e longitudinal dos dados. Para tornar o modelo mais robusto, foram comparadas duas propostas de inclusão da informação de similaridade química entre os compostos, utilizando o conceito de espaço químico representado por uma rede. O modelo obtido desta forma foi adaptado para servir em tarefas de ajuste de curvas Dose-Resposta a dados longitudinais; e *scripts* em *R* foram construídos com o intuito de permitir a aplicação automatizada futura destes modelos. O melhor modelo univariado proposto neste trabalho foi aquele que assume uma distribuição Poisson Generalizada, onde efeitos aleatórios a nível de verme foram necessários para a construção de um modelo adequado aos dados. Em busca de estender a modelagem anterior para uma análise fenotípica mais ampla, foi realizada uma análise multivariada para testar um modelo preditivo de compostos *hits*, do tipo árvore de regressão com efeitos aleatórios. Com base em uma análise exploratória inicial, foi feita a transformação e redução de dimensionalidade dos dados. Avaliou-se a possibilidade de discriminar fenótipos diferentes por meio de agrupamento *fuzzy* dos dados disponíveis. Entretanto, foi possível discriminar apenas dois *clusters* nesta etapa de agrupamento.

Ministério da Saúde

FIOCRUZ
Fundação Oswaldo Cruz

INSTITUTO OSWALDO CRUZ

DEVELOPMENT OF STATISTICAL MODELS FOR AUTOMATED PHENOTYPIC ASSAYS OF ADULT *SCHISTOSOMA MANSONI*

ABSTRACT

MASTER DISSERTATION IN BIOLOGIA COMPUTACIONAL E SISTEMAS

Fábio Jorge de Vasconcellos Júnior

Among the parasitic diseases, schistosomiasis is a serious disease that afflicts part of the poorest population. Therefore, the search for new pharmacological therapies against its etiologic agent (the parasites of the genus *Schistosoma*) is an activity of great social and economic relevance. However, a single drug, which already faces problems against resistant strains, is currently available in practice. The first step in the research and drug development process is the identification of compounds with therapeutic activity (*hits*), whose commonly used method is the screening of a library of compounds based on the target or based on phenotype. Against parasitic diseases, the phenotypic approach is very relevant, thanks to its greater biological value (systemic), compared to a biochemical test. In addition, as its modern extension, automated high-content screening (HCS) methods are available in cells or in small organisms. One of the difficulties of this approach lies in the insufficiency of the most widespread statistical tools in the analysis of more complex data. Therefore, it is necessary to develop models for a statistical analysis appropriate to the complex data obtained experimentally, in order to estimate the effect of the compound on the phenotype of the parasite. Thus, the objective of the present work was the development of statistical models to analyze the effect of antischistosomal candidate compounds from quantitative data from automated phenotypic assays obtained in adult *Schistosoma mansoni* with HCS technology in *Bioensaios e Triagem de Fármacos* platform of IOC/FIOCRUZ. From the exploratory statistical analysis of the data was possible to propose and compare models for the inference of the effect of each compound tested, from parasite motility data. Multilevel Generalized linear models were evaluated, in order to match the hierarchical and longitudinal structure of the data. In order to make the model more robust, two proposals were compared for the inclusion of chemical similarity information among the compounds, using the concept of chemical space represented by a network. The model obtained in this way was adapted to serve in tasks of adjustment of Dose-Response curves to longitudinal data; and R-scripts were built to allow the future automated application of these models. The best univariate model proposed in this work was one that assumes a Generalized Poisson distribution, in which random effects at the worm level were necessary to specify a suitable model to the data. In order to extend the previous modeling to a broader phenotypic analysis, a multivariate analysis was performed to test a regression tree predictive model with random effects to find hits compounds. Based on an initial exploratory analysis, the transformation and reduction of dimensionality of the data was done. The possibility of discriminating different phenotypes was evaluated through fuzzy clustering of the available data. However, it was possible to discriminate only two clusters in this clustering phase.

ÍNDICE

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE SIGLAS E ABREVIATURAS

BN	Binomial Negativa
GP	Poisson Generalizada (do inglês, *Generalized Poisson*)
HCA	Análise de alto conteúdo (do inglês, *High Content Analysis*)
HCS	Triagem de alto conteúdo (do inglês, *High Content Screening*)
IID	Independentes e identicamente distribuídos
INLA	Aproximação de Laplace Integrada e Aninhada (do Inglês, *Integrated Nested Laplace Approximation*)
GLM	Modelo Linear Generalizado (do inglês, *Generalized linear model*)
LGM	Modelos Gaussianos Latentes (do inglês *Latent Gaussian Models*)
log	função logaritmo natural
MGLM	Modelo Linear Generalizado Multinível (do inglês, *Multilevel Generalized linear model*)
ND	Não disponível
PCA	Análise de Componentes Principais (do inglês *Principal Component Analysis*)
PZQ	Praziquantel
RE-EM	Algoritmo Maximização de Expectativas com Efeitos aleatórios (do inglês, *Random Effect – Expectation-maximization algorithm*)
RW2	Modelo de passeio aleatório de segunda ordem (do inglês, *Second-order Random Walk model*)
SEM	Modelo de erro espacial (do inglês, *Spatial Error Model*)
SNG	Gerador de Rede de *Scaffolds* (do inglês, *Scaffolds Network Generator*)
sp.	espécie
WAIC	Critério de Informação largamente aplicável (do inglês, *Widely Applicable Information Criterion*)
ZINB	Binomial Negativa inflada por zeros (do inglês, *Zero-Inflated Negative Binomial*)

1 INTRODUÇÃO

A esquistossomose é uma doença negligenciada por estar ligada a condições de extrema pobreza, principalmente nos países em desenvolvimento, recebendo pouca atenção da indústria farmacêutica (CIOLI et al., 2014). Não obstante, a doença é responsável por aproximadamente 280.000 mortes por ano e alta morbidade em mais de 200 milhões de pessoas, o que afeta a qualidade de vida e a produtividade econômica dos contaminados (CAFFREY, 2015). Ela é uma helmintíase causada por vermes trematódeos do gênero *Schistosoma*, onde as três principais espécies patogênicas são a *S. mansoni*, *S. haematobium* e *S. japonicum*. No Brasil, a espécie S. *mansoni* é o agente etiológico desta infecção, conhecida popularmente como “xistose”, “barriga d’água” e “doença dos caramujos”. Ela é endêmica em diferentes estados do Nordeste do país e em Minas Gerais. A infecção crônica está associada a quadros debilitantes como anemia, nanismo e diminuída capacidade física e mental (HOTEZ et al., 2006).

1.1 O ciclo de vida do gênero *Schistosoma*

O ciclo evolutivo dos esquistossomos inclui dois hospedeiros obrigatórios: um hospedeiro definitivo no qual o parasito adulto se reproduz sexuadamente e um hospedeiro intermediário *Biomphalaria sp.*, molusco no qual o parasito se multiplica assexuadamente (ROSS et al., 2002). A transmissão do mamífero para o molusco é assegurada por uma larva de vida livre, o miracídio, que quando gerado a partir do ovo penetra ativamente no invertebrado. A transmissão do molusco para o mamífero é feita por outra larva de vida livre, a cercária, que também penetra ativamente na pele do hospedeiro mamífero. A cercária migra através da pele e atinge o sistema vascular, onde se transforma em esquistossômulo e em seguida no verme adulto, com dimorfismo sexual (MONÉ e BOISSIER, 2004). Os machos e fêmeas permanecem em cópula durante toda a sua vida, enquanto a fêmea libera centenas ou milhares de ovos diariamente. Eles podem ser excretados nas fezes ou aprisionados em tecidos adjacentes provocando reações granulomatosas que são responsáveis pela doença (COLLEY et al., 2014). Portanto, conhecendo-se o seu ciclo de vida, nota-se a importância de serem realizados estudos fenotípicos em

parasitas adultos, pois é neste estágio que os sintomas da doença se tornam visíveis (MONÉ e BOISSIER, 2004).

1.2 O tratamento, pesquisa e desenvolvimento farmacológico

Atualmente, o praziquantel (PZQ) é o único tratamento largamente disponível para a esquistossomose. Ele é um medicamento com comprovada atividade esquistossomicida e apresenta outras vantagens, como o fato de ser administrado ao paciente em dose única (CIOLI et al., 2014). Contudo, estudos indicam que cepas laboratoriais resistentes a PZQ podem ser isoladas e, além disso, foram relatadas ocorrências de isolados clínicos com menor susceptibilidade ao fármaco (WANG et al., 2012). Portanto, é uma questão de tempo até que a resistência se torne um problema de saúde pública. Ademais, PZQ é muito menos ativo contra vermes juvenis, o que muitas vezes resulta em curas incompletas (ARAGON et al., 2009). Seu mecanismo de ação, incluindo a sua biotransformação, não é totalmente compreendido (CIOLI et al., 2014), o que dificulta o desenvolvimento racional de variantes deste fármaco. Um número relativamente pequeno de derivados do PZQ foi sintetizado e testado, mas nenhum composto promissor com desempenho melhor que PZQ foi identificado e escassa informação pôde ser obtida de relações estrutura-atividade (CIOLI et al., 2014).

Temos como alternativa a esta abordagem, ou complementarmente a ela, a estratégia da busca por novos fármacos através da triagem de coleções de compostos químicos contra os parasitas, por meio de testes em modelos animais infectados (*in vivo*) ou por avaliação dos helmintos *in vitro* (PAVELEY e BICKLE, 2013). Justamente, uma das principais limitações para descobrir novos medicamentos anti-helmínticos é a ausência de um ensaio rápido, quantitativo e reprodutível para caracterizar a atividade das moléculas candidatas contra parasitas adultos (RAMIREZ et al., 2007). Associado a este método é necessário a aplicação de modelos apropriados para a análise estatística dos dados experimentais com vermes adultos. É esta lacuna que este trabalho tem como propósito preencher.

1.3 Revisão Bibliográfica

O paradigma de ensaio atualmente aceito para os vermes adultos envolve a anotação manual dos aspectos morfológicos e de motilidade do parasita com base na observação ao microscópio, que pode ter o resultado comprometido devido a falhas de observação e falta de embasamento estatístico. Como sua moderna extensão, temos as técnicas de análise e triagem de alto conteúdo (HCA e HCS, respectivamente, do inglês *High Content Analysis* e *High Content Screening*), proporcionando a captura automatizada de imagens com um instrumento de alta resolução. Esse método permite extrair informações detalhadas, com *softwares* de análise especializados, das imagens obtidas por microscopia automatizada, favorecendo uma melhor compreensão sobre as estruturas analisadas (ZANELA et al., 2010; CRUZ et al., 2013). A tecnologia HCS tem ganhado popularidade, especialmente pela indústria farmacêutica para conduzir ensaios fenotípicos baseados em células, para os quais já foram obtidos grandes avanços (ZANELLA et al, 2010).

A análise por microscopia automatizada de helmintos, em ensaios *in vitro* para determinar os efeitos de fármacos, é uma ferramenta precisa na classificação e reconhecimento de objetos de estudo em imagens digitais (BULLEN, 2008; CASTAÑÓN, 2006). A falta de linhagens transgênicas expressando proteínas fluorescentes em helmintos têm limitado os esforços experimentais no âmbito da microscopia de campo claro. De todo modo, esta técnica tem como vantagem em relação ao uso de marcadores fluorescentes exigir menos passos de manipulação e evita o uso de fluoróforos potencialmente tóxicos que podem afetar o resultado do ensaio (GRAVES, 2011). Adicionalmente, a microscopia de campo claro pode simultaneamente fornecer dados sobre muitos parâmetros fenotípicos, tais quais tamanho, forma, granularidade, vacuolização, danos tegumentais e a motilidade (PAVELEY e BICKLE, 2013).

O grande desafio associado às técnicas de análise fenotípica de helmintos por microscopia automatizada está no desenvolvimento de algoritmos customizados para quantificar esses efeitos de forma automática (PAVELEY et al., 2012). O uso de algoritmos para o processamento de imagens biológicas tem sido largamente aplicado em ensaios baseados na visualização de células (JONES et al, 2009; PERLMAN et al, 2004; TANAKA et al, 2005; MEGASON e FRASER, 2007;

STARKUVIENE e PEPPERKOK, 2007; LJOSA e CARPENTER, 2009; KIMMEL et al, 2017) e do nematódeo *Caenorhabditis elegans* (CRONIN et al, 2005; CONERY et al, 2014; TIAN et al, 2010; WAHLBY et al, 2012; GENG et al, 2004; ALBRECHT e BARGMANN, 2011). Como resultado, várias ferramentas de análise estão disponíveis para segmentação e análise a nível celular, por exemplo, o *software CellProfiler* (CARPENTER et al, 2006; LAPRECHT et al, 2007). Muitos estudos foram direcionados também para a análise do *C. elegans*, pois ele é um importante modelo para estudo em biologia. Encontra-se disponível para uso a ferramenta chamada "*WormToolbox*" também através do projeto *open-source CellProfiler*, permitindo a análise de ensaios com base em imagens de alto conteúdo de *C. elegans* para o estudo de diversas vias biológicas que são relevantes para doenças humanas (WAHLBY et al, 2012). Algumas estratégias para análise de fenótipos celulares incluem ajuste de modelos elípticos (BAI et al, 2009) e combinação estatística de modelos construídos manualmente (LIN et al, 2007). Para *C. elegans*, os métodos recentes incluem o uso de modelos articulados (HUANG et al, 2009), e modelos deformáveis do tipo "*Probabilistic shape models*" (WAHLBY et al, 2010), que, seguindo um critério de otimização, deforma um contorno definido para encontrar um objeto conhecido em uma determinada imagem (FISKER, 2000).

Entretanto, modelos explícitos do tamanho, forma e intensidade não são práticos para esquistossomos de qualquer estágio evolutivo devido à variação natural dos indivíduos (já que não existem clones de esquistossomos). Além disso, quando estes são expostos a situações de estresse, as mudanças na morfologia e aparência dos esquistossomos são muito mais complexas e dinâmicas que as mudanças fenotípicas observadas em *C. elegans* (ASARNOW e SINGH, 2013; SINGH, 2012).

No que diz respeito à análise dos dados obtidos após o processamento das bioimagens de helmintos, foi demonstrada (SINGH et al, 2009) a viabilidade de extração de informação e classificação automatizada de ensaios fenotípicos através de um processo de segmentação da imagem para identificação de parasitas individuais e representação do parasita no seu espaço de características e, por fim, classificação fenotípica usando-se árvores de classificação e regressão (CART – do inglês, "*Classification and regression trees*"). Desde então um rápido progresso tem sido feito no desenvolvimento e otimização de técnicas de análise fenotípica quantitativa e automatizada (de vermes em estágio larval). Dentre outros métodos

para detecção de compostos bioativos (PAVELEY et al, 2012; MARCELLINO et al, 2012), pode ser citado, por exemplo, o desenvolvimento de algoritmos para representação e comparação de respostas fenotípicas na forma de séries temporais (LEE et al, 2012). Naquele trabalho, foram apresentados algoritmos para agrupamento e análise de fenótipos incluindo um para quantificar a variabilidade fenotípica da população de parasitas analisada. Entretanto, nenhum método foi capaz de determinar relações entre dose/tempo e resposta fenotípica. O software QDREC, baseado no aprendizado supervisionado para identificar parasitas afetados por fármacos a uma dada concentração e tempo de contato, foi um avanço recente neste contexto (ASARNOW e SINGH, 2013). Mais recentemente, buscou-se também o desenvolvimento de algoritmos úteis para mapear e caracterizar o conjunto das respostas fenotípicas dos parasitas induzidas por fármacos (SINGH et al, 2016) de modo semelhante à modelagem do comportamento de *C. elegans* que já havia sido feita (ROUSSEL et al, 2007; STEPHENS et al, 2008). Sabe-se, porém, que os desafios no estudo do comportamento dos esquistossomos são mais complexos, vistas as limitações experimentais a eles associadas e citadas anteriormente.

Em sua maioria, os estudos de HCA/HCS aplicados à identificação de novos fármacos contra a esquistossomose têm feito uso do verme em estágio larval chamado esquistossômulo (PAVELEY et al., 2012; PAVELEY e BICKLE, 2013; ASARNOW et al., 2015), pois estes estão disponíveis em maior número que o parasita adulto, podendo ser produzidos *in vitro* com baixo custo e são pequenos o suficiente para serem usados facilmente em ensaios em microplacas, entre outras vantagens (PAVELEY e BICKLE, 2013). Entretanto, a realização de ensaios para avaliar a suscetibilidade de vermes adultos aos diferentes compostos químicos testados – apesar das limitações experimentais – são fundamentais para a descoberta de novos fármacos esquistossomicidas, pois é neste estágio que o parasita se apresenta quando os sintomas da doença evoluem (MONÉ e BOISSIER, 2004).

Em publicações recentes do nosso grupo, aplicou-se a técnica de HCS em parceria com outros métodos computacionais como a triagem virtual baseada em relações quantitativas estrutura-atividade, abreviada em inglês por *QSAR-Based Virtual Screening* (NEVES et al., 2016a; MELO-FILHO et al., 2016), e quimiogenômica aliada ao reposicionamento de fármacos (NEVES et al., 2016b), que renderam a identificação de promissores *hits* antiesquistossomais agora

disponíveis para otimização. Nesse trabalho, pretende-se complementar os métodos supracitados, desenvolvendo-se uma ferramenta automatizada de análise dos dados obtidos pela técnica de HCA para realizar a triagem de compostos químicos em vermes adultos de *S. mansoni*, e assim permitir a identificação de novos fármacos para a esquistossomose mansônica.

Com o processamento e análise das imagens do parasito podemos ter dados quantitativos ligados a várias características, como o tamanho, a forma e motilidade. Resultados anteriores do grupo demonstram a viabilidade do método de análise de imagens (NEVES et al., 2016a; NEVES et al., 2016b; MELO-FILHO et al., 2016). O desafio agora compreende a realização de análises estatísticas automatizadas aplicáveis aos dados de ensaios fenotípicos por HCA, buscando o aperfeiçoamento da plataforma de alto rendimento. Com este projeto, espera-se superar as dificuldades para a quantificação e a caracterização, estatisticamente robustas, do efeito dos fármacos em potencial sobre as formas adultas de *S. mansoni,* pelo desenvolvimento de modelos para análise dos dados derivados de experimentos de HCA. Têm-se como hipótese do trabalho que é possível e suficiente uma modelagem estatística ao nível de medida, que contemple a estrutura hierárquica e o caráter longitudinal dos experimentos. Adicionalmente, é esperado que a inclusão de informação de similaridade química entre os compostos - utilizando o conceito de espaço químico representado por uma rede (MAGGIORA e BAJORATH, 2014) – torne mais robusto o modelo proposto para análise de experimentos de triagem de compostos químicos.

1.4 Métodos de Processamento de Imagens

Antes do processamento de imagens biológicas propriamente dito é necessário um pré-processamento das imagens para eliminar artefatos produzidos pelo microscópio ou câmera usados na aquisição das imagens (SOMMER e GERLICH, 2013). Entre eles, pode-se citar a necessidade de corrigir a imagem para compensar, com o uso de imagens de referência, a iluminação desigual produzida pelo microscópio (BUCHSER et al., 2004).

A primeira etapa do processamento é a detecção de objetos de interesse, que podem ser, por exemplo, células ou organismos inteiros. Esta etapa é comumente chamada como a etapa de segmentação (SOMMER e GERLICH, 2013). Este

procedimento é realizado por meio de algum método que consiga distinguir entre o que é definido como um objeto e o plano de fundo da imagem. Pode ser feito, por exemplo, por reconhecimento de um conjunto de *pixels* em uma região da imagem com maior brilho (após definição pelo usuário de um limiar de intensidade de *pixels*), ou pode ser baseado na determinação primária do contorno dos objetos (a partir da medida de gradientes na intensidade de *pixels*, como uma mudança abrupta de intensidade entre *pixels* vizinhos).

Após a segmentação, é possível “mascarar” regiões que não sejam de interesse nos cálculos posteriores. A partir dos objetos e das imagens são obtidos valores de alguns tipos de medidas que formam a base das análises quantitativas subsequentes. Com o software *CellProfiler* (CARPENTER et al, 2006; LAPRECHT et al, 2007), é possível produzir medidas e estatísticas de vários tipos. Uma das grandes vantagens deste *software* é o seu *design* modular de *workflow*, com módulos diversos que agrupam cada um diversas medidas relacionadas. Em especial, pode-se citar *MeasureGranularity*, *MeasureTexture*, *CalculateImageOverlap* e *MeasureImageIntensity* (CARPENTER e JONES).

O módulo *MeasureGranularity* produz medidas de textura que busca ajustar uma série de elementos estruturais de tamanho crescente e retorna um espectro de medidas da qualidade deste ajuste à textura da imagem. O módulo *MeasureTexture* mede os descritores de Haralick (HARALICK et al, 1973) e uma estatística derivada de filtros de Gabor (GABOR, 1946) para cada objeto em uma imagem. Estas medidas também são determinadas pela textura da imagem, mas são resultado da comparação entre *pixels* vizinhos. Este método de análise textural consiste no cálculo de uma matriz de correlação de intensidade de *pixel*, entre os *pixels* comparados de um objeto ou imagem. Os descritores de Haralick compreendem 14 estatísticas diferentes derivadas da matriz de correlação obtida. Já os filtros de Garbor permitem detectar bandas correlacionadas de intensidade de *pixel*, por identificar frequências espaciais de oscilação na intensidade *dos pixels*, bem como a orientação espacial associada. O módulo *CalculateImageOverlap* calcula diversas estatísticas a partir de imagens preto e branco, a partir do grau de sobreposição entre regiões brancas de duas imagens. O módulo *MeasureImageIntensity* mede a intensidade total de uma imagem por somar todas as intensidades de pixels da imagem (exceto *pixels* mascarados); produz também estatísticas-resumo (média,

mediana, mínimo e máximo de intensidade) e número total de pixels, dada a mesma restrição.

1.5 Modelos Lineares Generalizados

Modelagem Linear Generalizada é uma abordagem que inclui a regressão linear e a regressão logística como casos particulares. Um modelo linear generalizado (GLM, do inglês, *Generalized linear model*) envolve (GELMAN e HILL, 2007):

1. Um vetor de dados resposta $y = (y_1, y_2, ..., y_n)$;

2. Uma matriz de preditores $\boldsymbol{X}$ e um vetor de coeficientes β, que formam um preditor linear $\eta = \boldsymbol{X}\beta$;

3. Uma função *link* g, tal que a sua função inversa g^{-1} aplicada ao preditor linear resulte em um vetor de dados transformados $\omega = g^{-1}(\eta)$ que são usados para modelar os dados y.

4. Uma distribuição de probabilidade condicionada a ω, $p(Y|\omega)$, pertencente à família exponencial, atribuída à variável aleatória Y que tem os dados y como uma amostra aleatória.

5. Possivelmente outros parâmetros, formando um novo vetor θ, que estão associados ao preditor linear η, à função *link* ou diretamente à distribuição. Desta forma, a distribuição em (4) pode ser reescrita em termos de θ, como $p(Y|\omega,\theta)$ e será chamada função verossimilhança (ou *Likelihood*, em inglês), quando escrita como uma função dos parâmetros (ω,θ).

Diferentes GLMs devem ser citados aqui. A regressão linear é o caso onde a função *link* é simplesmente a função identidade $g(u) = u$, a distribuição é gaussiana e o desvio padrão σ é estimado a partir dos dados.

Em especial, é relevante citar as GLMs aplicáveis a dados y do tipo contagem, isto é, y_i inteiro não-negativo. Enquanto que no estudo de variáveis contínuas a distribuição gaussiana é a referência, para variáveis contáveis a distribuição de Poisson é geralmente o ponto de partida. A distribuição de Poisson é um modelo simples, descrito por um único parâmetro λ, que é simultaneamente a média e a variância. Geralmente, em um GLM Poisson, λ é tomado como $\lambda = g^{-1}(\boldsymbol{X}\beta)$, onde g é a função logarítmica natural. Ela expressa a probabilidade de um dado número de eventos ocorrer em um intervalo de tempo fixo unitário e que é independente do

número de eventos que ocorreram antes deste intervalo. Além disso, a taxa de ocorrência λ é constante e não ocorrem dois ou mais eventos simultaneamente (JAMES, 2015).

1.6 Distribuições de probabilidade para dados de contagem

Como é comum que as restrições associadas a uma regressão Poisson não sejam suposições realistas sobre os dados a se modelar, uma alternativa é testar distribuições que sejam reconhecidas generalizações da distribuição de Poisson, mesmo sendo mantida a função *log-link*. A distribuição Binomial Negativa (BN) é um exemplo notório e muito usado na análise de dados de contagem, pois ela possui um parâmetro extra, chamado "parâmetro de dispersão", que a permite ajustar dados que tenham variância maior que a média (GELMAN e HILL, 2007); ou também conhecido como "parâmetro de *clustering*" (LLOYD-SMITH, 2007), pois é útil na modelagem de contagens onde os "eventos" ocorrem em grupos. Ela também é chamada de distribuição Gama-Poisson, pois ela é equivalente a uma distribuição Poisson em que a taxa de ocorrência λ é, em si, uma variável aleatória que segue uma distribuição gama (JAMES, 2015). Pode-se demonstrar que a distribuição BN está relacionada a processos de ocorrência de eventos "sucesso" ou "fracasso" com probabilidades *p* e (1 - *p*), respectivamente. A função massa de probabilidade *P(y)* de uma variável aleatória Y distribuída segundo uma binomial negativa é

$$\mathrm{P}(y) = \frac{\Gamma(y+n)}{\Gamma(n)\Gamma(y+1)} p^n (1-p)^y \ , \quad y = 0, 1, 2, \ldots$$

onde *P(y)* abrevia *P(Y=y)*, *n* é o parâmetro de dispersão, Γ é a função gama (JAMES, 2015). A média μ e a variância σ^2 de Y são:

$$\mu = n\frac{1-p}{p} \qquad \mathrm{e} \qquad \sigma^2 = \mu(1 + \frac{\mu}{n})$$

Nota-se que, diferente da distribuição Poisson, a sua variância é uma função quadrática da média para a distribuição BN.

Outra distribuição relevante para dados de contagem é a distribuição chamada Poisson Generalizada GP (CONSUL e JAIN, 1973). Em especial, seu emprego em modelagem estatística é atualmente dado por meio da parametrização

chamada GP-P (ZAMANI e ISMAIL, 2012). A função de probabilidade *P(y)* desta distribuição Poisson Generalizada na forma GP-P é:

$$\mathrm{P}(y) = \frac{\mu(\mu+\varphi\mu^{p-1}y)^{y-1}}{(1+\varphi\mu^{p-1})^{y}y!}\exp\left(-\frac{\mu+\varphi\mu^{p-1}y}{1+\varphi\mu^{p-1}}\right) \quad , \quad y=0,1,2,\ldots$$

onde μ é o parâmetro média; φ é o parâmetro de dispersão e p é o parâmetro de forma, que pode ser fixado *a priori*. A variância σ^2 de Y é uma função da média e dos demais parâmetros:

$$\sigma^2 = \mu\left(1+\varphi\mu^{p-1}\right)^2$$

Note que o parâmetro *p* é quem determina a ordem da relação entre variância e média. Por exemplo, se o parâmetro *p* for tomado igual a 1, a variância será proporcional à média; mas, se p = 1,5 então esta relação é de segunda ordem.

Outras distribuições que podem ser interessantes para dados de contagem são as já bem conhecidas distribuição binomial e sua generalização chamada beta-binomial. A distribuição binomial é classicamente reconhecida como a lei que prevê a probabilidade, por exemplo, de ocorrer Y “caras” em *n* lançamentos de uma moeda onde a probabilidade *p* de sair “cara” é constante; enquanto a beta-binomial estende este conceito para a probabilidade de ocorrer eventos “bem-sucedidos” (como sair “cara”, no exemplo anterior) onde a probabilidade *p* é variável, dada por uma distribuição beta (JAMES, 2015). Elas têm a característica de assumirem valores de Y somente dentro de uma faixa de zero a *n*. Isto pode ser útil na modelagem de dados naturalmente limitados em extensão (GELMAN e HILL, 2007). A função de probabilidade *P(y)* da distribuição beta-binomial é

$$\mathrm{P}(y) = \binom{n}{y}\frac{B(y+\alpha, n-y+\beta)}{B(\alpha,\beta)} \quad , \quad y=0,1,\ldots,n.$$

onde *B* é a função beta (JAMES, 2015); os parâmetros α e β são os parâmetros de forma da distribuição beta geradora da distribuição beta-binomial. A média μ e a variância σ^2 de Y

$$\mu = n\frac{\alpha}{\alpha+\beta} = n\mu_p \qquad \text{e} \qquad \sigma^2 = n\mu_p(1-\mu_p)\left(1+(n-1)\rho\right)$$

onde ρ é identificado como um parâmetro de correlação ($0 < \rho < 1$), dado por $\rho = (\alpha + \beta + 1)^{-1}$; μ_p é a média da probabilidade *p*, citada anteriormente.

Há casos onde uma distribuição de frequência dos dados reais indica que a ocorrência de valores zero é muito maior que o esperado. Neste caso, modificações das distribuições anteriores, considerando este excesso de zeros é uma opção para a modelagem dos dados em estudo. Existem diversos modelos para este fim, chamados modelos inflados por zeros, em diversas aplicações (MULLAHY, 1986; LAMBERT, 1992; DALRYMPLE, et al, 2003; FREES, 2011). A função de probabilidade *P(y)* desta extensão (exemplificando para a distribuição Poisson) é

$$\mathrm{P}(y) = p \times 1_{[y=0]} + (1-p) \times \mathrm{Poisson}(y) \ , \qquad y = 0, 1, 2, \ldots$$

onde *p* é a probabilidade de ocorrer zeros extras e $1_{[y=0]}$ representa a função indicadora de y = 0. Um modelo[1] particularmente interessante para *p* é um que assume a relação

$$p = 1 - \left(\frac{\exp(\eta')}{1 + \exp(\eta')} \right)^{\alpha}$$

onde *η'* é o preditor linear do modelo (somado a um possível *offset*[2]), e α é um parâmetro que governa o grau de inflação por zeros (LAMBERT, 1992). Esta modelagem é útil quando for teorizado que a probabilidade de resultados nulos aumente com a redução do preditor linear.

1.7 Inferência Bayesiana, modelos multinível e INLA

Em uma modelagem estatística, é necessário que as características de interesse serem inferidas a partir dos dados disponíveis; isto é, em um GLM, por exemplo, é preciso aplicar alguma abordagem para inferir os coeficientes *β* e os parâmetros *θ*. Já que estes são desconhecidos, ao estimá-los é importante conhecer também o grau de incerteza que permanece sobre cada uma destas estimativas. Uma forma de atender a estes interesses é adotar a inferência bayesiana, que tem como diferencial uma relativa simplicidade de interpretação e manipulação das estimativas e suas incertezas. Nesta abordagem, os coeficientes e parâmetros do

[1] Neste trabalho, este modelo foi o único tipo de modelo inflado por zeros usado. Assim, por simplicidade, as futuras referências a modelos inflados por zero presentes nesta dissertação se referem a este modelo especificamente.

[2] *Offset* é um termo utilizado em modelagem estatística para se referir a uma variável com coeficiente *β* conhecido (GELMAN e HILL, 2007).

modelo são também tratados como realizações de variáveis aleatórias, assim como é feito de antemão para os dados *y* (GAMERMAN e LOPES, 2006). Na inferência bayesiana, informações *a priori* a respeito dos *coeficientes* e *parâmetros*[3] são fornecidas na forma de uma distribuição de probabilidade, $p(\omega,\theta)$. Suportada pelo teorema de Bayes, pode-se obter a distribuição *a posteriori* de (ω,θ) condicional aos dados - $p(\omega,\theta|y)$ - graças a combinação entre a distribuição dos dados $p(Y|\omega,\theta)$ e a distribuição *a priori* $p(\omega,\theta)$. As estatísticas de interesse são as estatísticas resumo da distribuição *a posteriori* (GAMERMAN e LOPES, 2006).

Apesar de os modelos multinível não serem obrigatoriamente resultados da aplicação de métodos de inferência bayesiana, sua descrição é mais simples a partir desta ótica. Em especial, modelos GLM multinível (MGLM, do inglês *Multilevel Generalized linear model*) são uma extensão de GLM, onde os *coeficientes β* em si são modelados por uma distribuição a priori $p(\beta|\theta')$, onde θ' é um vetor adicional de parâmetros desconhecidos (chamados hiperparâmetros), o qual tem também, por sua vez, uma distribuição a priori, $p(\theta')$, chamada hiperpriori. É esta construção em camadas (ou níveis) que justifica o nome modelos multinível, ou também conhecidos como modelos hierárquicos. Nesta configuração, os *coeficientes β* são chamados efeitos aleatórios (GELMAN e HILL, 2007).

MGLMs são modelos estatísticos adequados para descrever um conjunto de dados inerentemente estruturados em diferentes níveis ou grupos, pois permitem a modelagem simultânea da variação entre grupos e da variação interna de cada grupo (variação local), usando todos os dados disponíveis (GELMAN e HILL, 2007). Um dos benefícios do uso desta abordagem multinível advém da adequação deste tipo de modelagem na produção de inferências para grupos com pequeno tamanho de amostra, que é o caso dos dados aqui analisados neste trabalho. Por um lado, sabe-se que a aplicação da estimativa clássica, que usa apenas informação local, pode ser essencialmente inútil se o tamanho da amostra for pequeno, enquanto que a regressão clássica usando todos os dados – mas ignorando os indicadores de grupo – pode ser enganosa. Entre estes dois extremos, existe a modelagem hierárquica, que permite estimar médias de grupo e efeitos em nível de grupo, comprometendo-se entre a estimativa excessivamente influenciada pelo ruído dentro

[3] Os termos "*parâmetros*" e "*coeficientes*" serão mantidos em uso, por puro abuso de linguagem. Em geral, estes termos perdem o sentido original empregado na abordagem frequentista (GAMERMAN e LOPES, 2006).

do grupo e a estimativa de regressão simplificada que ignora os indicadores de grupo (GELMAN e HILL, 2007).

A inferência bayesiana é normalmente implementada usando métodos computacionais, cujo principal método chama-se Monte Carlo via cadeias de Markov (MCMC, do inglês *Markov Chain Monte Carlo*), que utiliza simulações estocásticas para obter distribuições de probabilidade e tem larga aplicação (GAMERMAN e LOPES, 2006). Como os métodos via MCMC são computacionalmente custosos, é interessante usar uma metodologia alternativa baseada em aproximações de Laplace chamada INLA (do inglês, *Integrated Nested Laplace Approximation*), que fornece soluções rápidas e equivalentes ao MCMC para uma classe grande de modelos (RUE *et al.*, 2009). É uma metodologia aplicável à Modelos Gaussianos Latentes (LGM, do inglês *Latent Gaussian Models*), sendo muitos MGLMs casos particulares destes. Ele é obtido usando uma formulação hierárquica em 3 estágios,

$$\mathbf{y}|\mathbf{x}, \theta_2 \sim \prod_i p(y_i|\eta_i, \theta_2)$$

$$\mathbf{x}|\theta_1 \sim p(\mathbf{x}|\theta_1) = \mathcal{N}(0, \mathbf{\Sigma})$$

$$\theta = [\theta_1, \theta_2]^T \sim p(\theta)$$

onde as observações *y* são assumidas condicionalmente independentes, dado um campo aleatório latente gaussiano ***x*** e hiperparâmetros *θ*. O campo latente gaussiano nada mais é que uma especificação p(x|θ_1) assumida normal multivariada como única restrição. Os MGLMs são incorporados nesta formulação, tomando-se o preditor linear *η* como um somatório de efeitos fixos e termos *f* (rotulados como *componentes do modelo*) que são usados para especificar processos gaussianos específicos (RUE *et al.*, 2009).

Alguns exemplos de *componentes do modelo* (também chamados de *modelo latente*) são modelos de efeitos aleatórios com diferentes tipos de correlação, modelos espaciais e modelos para *smoothing*, como os bem conhecidos passeio aleatórios de primeira e segunda ordem (“RW1” e “RW2”, respectivamente, a partir do termo em inglês *random walk*). Componentes dignos de nota usados neste trabalho (além dos modelos RW1 e RW2) são:

a. Modelo IID para coeficientes **i**ndependentes e **i**denticamente **d**istribuídos;

b. Modelo condicional auto-regressivo chamado BYM (BESAG, YORK E MOLLIÉ, 1991), largamente aplicado na modelagem de efeitos aleatórios espaciais. A dependência espacial é expressa condicionalmente por requerer que o efeito aleatório em uma dada área, dados os valores das outras áreas, dependa somente dos valores em um pequeno conjunto de áreas vizinhas (BESAG, YORK E MOLLIÉ, 1991). A cada efeito espacial é somado um efeito não-espacial IID;

c. Modelo de Erro Espacial (SEM, do inglês *Spatial Error Model*), que contém um termo de erro autoregressivo espacial *u*, com um parâmetro de autocorrelação espacial ρ_{Err} associado a ele (BIVAND et al, 2014), dado pelas equações abaixo, exemplificando-se para o caso de uma regressão linear:

$$y = X\beta + u \quad ; \quad u = \rho_{\mathrm{Err}} W u + e \quad ; \quad e \sim MVN(0, \sigma^2 I_n)$$

onde *W* é uma matriz de pesos n × n que codifica as relações de dependência entre os erros u_i; *MVN* indica a distribuição normal multivariada e I_n é a matriz identidade de dimensão *n* (BIVAND et al, 2014). Este tipo de modelo é dito introduzir um termo de erro *espacial* por ser a sua aplicação mais difundida a modelagem de dados geográficos; ele pode, entretanto, ser utilizado para outros tipos de dependência entre os erros, pois a estrutura de dependência é totalmente definida pela matriz de pesos W, não importando sua origem.

d. Modelos que implementam efeitos não-lineares de variáveis preditoras, como o modelo chamado "*log1exp"* (LINDGREN e RUE, 2015). Ele permite construir um componente do modelo

$$\theta \log\left(1 + \exp(\alpha - \gamma x)\right)$$

em função de uma variável preditora positiva *x*, onde *θ*, *α*, *γ* são hiperparâmetros do modelo.

e. Modelo do processo estocástico *Ornstein-Uhlenbeck* (UHLENBECK e ORNSTEIN, 1930), que foi originalmente proposto para descrever a velocidade de uma partícula em movimento browniano, sob a influência do atrito. Seja x = (x_1, x_2, ..., x_n) os valores obtidos neste processo para instantes de tempo consecutivos *t* = (t_1, t_2, ..., t_n); para *i* = (2,3, ..., n) a distribuição condicional de x_i dados (x_1, ..., $x_{(i-1)}$) é gaussiana com média

$$x_{i-1}\exp(-\phi\delta_i)$$

e precisão (isto é, o inverso da variância)

$$\tau\left(1-\exp(-2\phi\delta_i)\right)^{-1}$$

onde δ_i é o intervalo de tempo entre o instante *i* e seu o instante anterior (*i* - 1); o parâmetro ϕ é um hiperparâmetro positivo; o hiperparâmetro τ é a precisão marginal do processo *Ornstein-Uhlenbeck.* Note que, tomando-se um valor fixo grande o suficiente para o hiperparâmetro τ (e x_1 fixo associado ao instante de tempo $t_1 = 0$), o modelo torna-se um efeito não-linear (e essencialmente não-estocástico) da forma

$$x_i = x_1\exp(-\phi t_i)$$

em função da variável preditora t_i, tendo hiperparâmetro desconhecido ϕ.

1.8 Representação em Rede do Espaço Químico

Espaço Químico é um conceito com origem na quimioinformática (que é uma área interdisciplinar, consistindo em técnicas computacionais aplicadas à problemas no campo da Química). Ele é geralmente concebido como numerosos grupos de compostos “espalhados” em um espaço multidimensional - de forma análoga às estrelas e demais corpos celestes agrupados em galáxias que habitam nosso universo - e tem grande aplicação no estudo da relação estrutura-atividade biológica de pequenos compostos, com impacto no desenvolvimento de fármacos (DOBSON, 2004).

Cada composto químico é um ponto deste espaço e a distância entre dois pontos quaisquer é definida em termos das suas propriedades fisico-químicas ou dissimilaridades estruturais. A representação do espaço químico que foi geralmente empregada é uma representação baseada em coordenadas, onde *n* propriedades são analisadas para cada composto químico, e estes são vistos cada um como um vetor contido em um espaço *n*-dimensional. Este tipo de representação sofre de alguns problemas, como, por exemplo, a alta dimensionalidade do espaço estudado, que dificulta a análise (MAGGIORA e BAJORATH, 2014).

Como alternativa, há também a possibilidade de se utilizar representações baseada em rede, como já se tem feito rotineiramente na representação e modelagem de redes sociais, redes de citação, entre outras (BARABÁSI, 2003).

Neste tipo de representação, os compostos são vértices conectados por arestas, em uma estrutura matemática grandemente estudada e chamada *grafo*. É o critério usado para gerar a conectividade entre os vértices que define a estrutura do espaço químico estudado. Tentativas de representar o espaço químico "biologicamente relevante" na forma de redes têm sido realizadas, com vantagens na visualização e análise dos dados (MAGGIORA e BAJORATH, 2014).

Uma forma de representar o espaço químico como uma rede é associar a cada par de compostos um coeficiente de similaridade química entre eles, dentre os quais o mais utilizado é o índice de Tanimoto, que apresenta vantagens quando comparado ao uso de métodos baseados em coordenadas somente (CHEN e REYNOLDS, 2002). A rede assim gerada é uma rede ponderada pelo grau de similaridade entre os compostos, e não dirigida[4]. Um outro tipo de representação em rede adotada para retratar o espaço químico é chamada de *Scaffolds Network* (VARIN et al, 2011; MATLOCK et al, 2013). Nela, os pequenos compostos orgânicos estudados têm identificadas suas subestruturas mais ou menos rígidas (dadas por conjuntos de cadeias cíclicas de carbono e heteroátomos), chamadas genericamente de *scaffolds*. A estrutura desta rede é gerada por associar compostos que possuam um mesmo *scaffold*, e conectar estes últimos dada a existência de *sub-scaffolds* em comum, isto é, substruturas também cíclicas, porém menores, presentes nestes *scaffolds*[5]. Assim, trata-se de uma rede tripartida dirigida formada por vértices identificados pelos compostos, seus *scaffolds* e *sub-scaffolds*. Esta abordagem pode ser muito relevante na área de descoberta de fármacos e química medicinal, onde *scaffolds* são muito utilizados no *design* e na racionalização dos resultados de pesquisa (VARIN et al, 2011).

1.9 Métodos de aprendizado estatístico

Aprendizado estatístico refere-se a um conjunto de ferramentas para modelar e entender dados complexos. É uma área de desenvolvimento recente em

[4] No contexto da teoria de redes complexas, uma rede pode ter arestas *dirigidas* ou *não-dirigidas*. Uma aresta dirigida é aquela que tem um sentido entre os dois vértices que ela une (representado por uma seta).

[5] Relativo a cada *scaffold* presente na rede, os seus *sub-scaffolds* podem ser classificados hierarquicamente em *sub-scaffolds* primários, secundários, etc. Os *sub-scaffolds* primários são as subestruturas contidas nos *scaffolds*, mas que não estão contidas em outros *sub-scaffolds*. Os *sub-scaffolds* secundários são, por sua vez, as subestruturas contidas nos *scaffolds e em, ao menos, um sub-scaffold primário.*

estatística e se mescla com desenvolvimentos paralelos em ciência da computação. O campo inclui muitos métodos, tal qual regressão "*lasso*", árvores de regressão e classificação e máquinas de vetores de suporte (JAMES et al, 2013).

Estas ferramentas podem ser classificadas como supervisionadas ou não-supervisionadas. Em resumo, aprendizado estatístico supervisionado envolve construir um modelo estatístico para predizer, ou estimar, uma variável-resposta (ou dependente), com base em uma ou mais variáveis preditoras (também chamadas variáveis independentes). Para a construção do modelo, parte-se de dados contendo valores tanto da variável-resposta quanto dos preditores. Diferentemente, uma aprendizagem estatística não-supervisionada utiliza dados das variáveis independentes e há ausência de variáveis-resposta para "supervisionar" a modelagem. Com métodos deste último tipo, pode-se aprender sobre relações entre variáveis ou sobre a estrutura dos dados estudados (JAMES et al, 2013).

Em geral, na criação de um modelo preditivo a partir de métodos de aprendizado estatístico, o conjunto de dados disponível é repartido aleatoriamente em um conjunto de dados principal (chamado de conjunto, ou grupo, *de treinamento*) e um conjunto de dados secundário, chamado de conjunto, ou grupo, *de teste*. O grupo de treinamento, como o nome indica, tem a função de "treinar" o modelo; isto é, serve como *input* para o algoritmo ou cálculo empregados na descoberta de parâmetros do modelo e demais *outputs* do processo. O grupo de teste, por sua vez, é aplicado ao modelo *já construído* a partir dos dados de treinamento, para poder avaliar a sua capacidade de previsão para dados "novos" ou "externos" (ao dados de treinamento), que é, de fato, o grande objetivo de um modelo preditivo (JAMES et al, 2013).

1.10 Análise de Componentes Principais

Análise de Componentes Principais (PCA, do inglês *Principal Component Analysis*) é uma abordagem popular para derivar um conjunto menor, mas representativo, de novas variáveis a partir de um grande conjunto de variáveis originais, mantendo-se grande parte da informação original. Em resumo, isto é feito realizando-se uma mudança de base e desprezando-se alguns vetores deste conjunto com menos importância na explicação da variabilidade dos dados.

Quando os dados originais são descritos por muitas variáveis correlacionadas, as novas variáveis, chamadas *componentes principais*, são encontradas sequencialmente ao longo das direções de maior variabilidade dos dados originais. O primeiro componente é determinado pela direção de maior variabilidade geral dos dados; o segundo componente é obtido como a direção ortogonal ao primeiro componente, de maior variabilidade não explicada pelo primeiro componente; o terceiro, por sua vez, é ortogonal ao dois primeiros componentes; e assim por diante para os componentes de ordem superior (JAMES et al, 2013). Em termos técnicos, PCA é uma transformação linear ortogonal que transforma os dados originais para um novo sistema de coordenadas, onde os componentes principais são autovetores da matriz X^TX (sendo X a matriz de dados originais com *n* observações × *p* variáveis) e a variância dos componentes são seus autovalores.

Um componente principal é uma combinação linear normalizada das variáveis originais, onde os coeficientes lineares são chamados *cargas*. Uma função das variáveis originais, chamada *função score,* pode ser obtida para cada componente principal, sendo então capaz de descrever uma observação qualquer nas novas coordenadas (JAMES et al, 2013).

Apesar da aplicação da PCA não ter como pré-requisito que os dados originais sejam oriundos de alguma distribuição aleatória pré-definida, como uma normal multivariada, é útil buscar reescalar e transformar os dados originais, caso estes assumam distribuições muito assimétricas. Isto é feito buscando obter distribuições mais simétricas, a fim de que a variância (a ser maximizada para cada componente principal) seja um parâmetro mais representativo da distribuição dos dados (JAMES et al, 2013).

1.11 Análise de Agrupamento

As técnicas de agrupamento têm como objetivo encontrar subgrupos do conjunto de dados por meio de um processo de aprendizado não-supervisionado (isto é, não há uma variável-resposta conhecida *a priori*). Os diferentes métodos buscam uma partição dos dados em grupos distintos que maximize a similaridade dentro do grupo, enquanto observações em grupos diferentes sejam bem distintas uma das outras. Para isso, é necessária a definição de uma medida da similaridade

ou, alternativamente, da distância entre um par de observações distintas (JAMES et al, 2013).

Um método de agrupamento para particionar um conjunto de dados em K subgrupos disjuntos, muito utilizado pela sua simplicidade e elegância, é conhecido na língua inglesa como *K-Means Clustering*. Neste método, após estar definida uma medida de distância adequada e o número K de *clusters*, computa-se K *centróides* que atuam cada um como uma média das observações pertencentes ao grupo ao qual é referência (JAMES et al, 2013). Cada observação em conjunto de dados é atribuída a um único grupo, e assim é o resultado da aplicação do *K-Means Clustering* e outros semelhantes. A fim de agrupar dados experimentais de forma mais versátil, existe uma generalização chamada *Fuzzy Clustering*, que estima graus de pertencimento (ou pertinência) de cada observação aos diferentes grupos detectados. O grau de pertencimento de uma observação a um determinado grupo *k* é representado por um coeficiente, com valor no intervalo [0, 1], chamado coeficiente de pertencimento ao grupo *k*, onde a soma dos coeficientes de cada observação é igual 1 (KAUFMAN e ROUSSEEUW, 2009). Observações são consideradas pertencer parcialmente a cada um dos grupos com graus de pertencimento diferentes; mas, por um abuso de linguagem, uma observação é também dita *pertencer* a grupo X, por exemplo, caso o seu coeficiente de pertencimento ao grupo X seja o maior entre os grupos. Uma generalização *fuzzy* disponível para o método *K-Means* se chama *Fuzzy C-Means*, que pode ter os coeficientes de pertencimento estimados utilizando o algoritmo FANNY (KAUFMAN e ROUSSEEUW, 2009). Este algoritmo tenta minimizar a função objetivo

$$\sum_{v=1}^{k} \frac{\sum_{i=1}^{n} \sum_{j=1}^{n} u_{iv}^{r} u_{jv}^{r} d(i,j)}{2 \sum_{j=1}^{n} u_{jv}^{r}}$$

onde u_{iv} é o coeficiente de pertencimento da observação *i* ao grupo *v*; o expoente *r* é um parâmetro de ajuste chamado, em inglês, *membership exponent*; k é o número de grupos; *n* é o número de observações; e *d*(*i*, *j*) é a distância entre uma observação *i* e outra *j*. A função objetivo é uma função das distâncias *d* e dos parâmetros *r* e *k*, e ela retorna os coeficientes de pertencimento, que conjuntamente a minimizam (MAECHLER et al, 2018). O expoente *r* é estritamente maior que a unidade, e ele controla o grau de sobreposição entre os *clusters* (quanto maior *r* for, mais "*fuzzy*" serão os *clusters*).

Daqui para frente, os grupos de observações identificados em uma análise de agrupamento serão chamados *clusters*, ao invés de “grupos” para distinguir entre outros tipos de grupos presentes neste trabalho.

1.12 Árvores de Regressão

Árvores de regressão são um caso particular de um tipo de modelo mais geral chamado Árvores de Decisão. A construção de uma árvore de decisão é um método de aprendizado supervisionado, que envolve a segmentação do espaço das variáveis preditoras X_1, X_2, ... , X_p em um número J de regiões disjuntas retangulares, de dimensão *p*, R_1, R_2, ... , R_J. Uma *árvore de regressão* é gerada quando a variável-resposta é contínua, enquanto ela é chamada *árvore de classificação* no caso de a variável-resposta ser categórica. A fim de fazer predições a partir de uma observação, identifica-se inicialmente em qual região R_j a observação está localizada e atribui-se a ela geralmente o valor médio $\hat{\mathbf{y}}_{R_j}$ da variável-resposta (caso esta seja contínua) – ou o valor modal, caso ela seja categórica – entre os dados de treinamento presentes na mesma região R_j da observação de interesse (JAMES et al, 2013). A construção de uma árvore de regressão é realizada pela busca das regiões R_j tal que seja minimizada a função objetivo, dada pelo somatório dos resíduos quadráticos

$$\sum_{j=1}^{J} \sum_{i \in R_j} (y_i - \hat{y}_{R_j})^2$$

Como é computacionalmente inviável considerar toda e qualquer partição do espaço das variáveis preditoras, a abordagem comumente usada para encontrar as regiões R_1, R_2, ... , R_J é conhecida como *recursive binary splitting*, ou divisão binária recursiva (JAMES et al, 2013). Neste método, primeiramente é selecionada uma variável X_j e um ponto de corte s_1 tal que o espaço das variáveis preditoras é dividido em duas regiões: $\{X_j < s_1\}$ e $\{X_j \geq s_1\}$. A variável X_j e o ponto de corte s_1 são aqueles que minimizam os resíduos quadráticos nesta primeira etapa. Em seguida, cada região obtida na primeira etapa é dividida em outras duas regiões, escolhendo-se outras variáveis X_k e X_l e um ponto de corte para cada variável, s_2 e s_3, (tal que também minimizem os resíduos quadráticos nesta etapa). O processo continua até que a função objetivo tenha convergido, ou até que outro critério de parada tenha

sido cumprido (JAMES et al, 2013). O resultado desta abordagem é um diagrama lógico em forma de árvore (isto é, uma estrutura ramificada) onde cada nó identifica uma regra de decisão (definida pela variável e valor de corte associados), que aponta um valor predito para cada região (JAMES et al, 2013).

O processo descrito acima pode produzir boas predições no conjunto de treinamento, mas provavelmente sobreajusta os dados, levando a um desempenho ruim no conjunto de teste. Para evitar o sobreajuste dos dados de treinamento, uma abordagem viável é “podar” a árvore de regressão (técnica conhecida em inglês pelo nome *Tree Pruning*). Em resumo, neste método busca-se encontrar uma sub-árvore a partir da árvore de regressão original que a supere em desempenho para a predição dos dados do conjunto de teste (JAMES et al, 2013).

Um tipo de modelo interessante que generaliza uma árvore de regressão é obtido pelo método, chamado *RE-EM tree*, que combina a estrutura de modelos de regressão linear multinível com a flexibilidade de um modelo baseado em árvore (SELA e SIMONOFF, 2012). Este método permite o ajuste de dados com medidas repetidas de um mesmo indivíduo, por exemplo (como em estudos com estrutura longitudinal), fornecendo estimativas para os efeitos aleatórios a nível de indivíduo. O modelo “híbrido”, obtido desta forma, relaciona a variável resposta Y com as variáveis preditoras (incluindo os metadados a nível de grupo) como apresentado pela equação

$$y_{it} = Z_{it}\mathbf{b}_i + f(x_{it1}, \ldots, x_{itK}) + \varepsilon_{it}$$

onde ε_{it} representa o erro residual associado a *t*-ésima observação de um indivíduo *i*; $\mathbf{b}_i$ é o vetor de efeitos aleatórios estimados para o indivíduo *i*, que segue uma distribuição normal centrada em zero e com desvio-padrão σ_{RE} ; Z_{it} é uma matriz de *design* conhecido que pode ser adicionada para incluir efeitos conhecidos *a priori* sobre os efeitos a nivel de indivíduo; e a função f(x_{it1}, ..., x_{itK}) representa a parte árvore de regressão do modelo em função das variáveis preditoras x_{it} (SELA e SIMONOFF, 2012).

2 OBJETIVOS

2.1 Objetivo Geral

Propor e avaliar modelos adequados aos dados oriundos do processamento de imagens de parasitas adultos, voltado para a identificação de novos fármacos esquistossomicidas.

2.2 Objetivos Específicos

1) Realizar análises exploratórias das variáveis fenotípicas para fundamentar a escolha das variáveis e a construção dos modelos;

2) Construir e comparar modelos multinível para a análise da motilidade do parasita (como indicador da sua vitalidade), buscando discriminá-la dos demais efeitos que contribuem para o valor final da resposta observada experimentalmente antes e após tratamento (contemplando experimentos do tipo triagem de compostos e experimentos do tipo Dose Resposta).

3) Construir um modelo para classificação de compostos como bioativos ou não-bioativos, com o uso de métodos de aprendizado estatístico.

4) Produzir *scripts* para a aplicação automática destes modelos em pesquisas futuras.

3 MATERIAL E MÉTODOS

3.1 Aquisição e processamento das bioimagens

Para a realização dos experimentos de microscopia, uma coleção de parasitas S. *mansoni* adultos foi obtida (usando procedimento apresentado na referência MELO-FILHO *et al*., 2016). Durante todo o experimento, vermes machos e fêmeas foram mantidos separados em placas de 96 poços, um indivíduo por poço, em meio DMEM completo e em temperatura e atmosfera controladas (37°C e 5% de CO_2). O estudo foi longitudinal, composto por 5 leituras feitas em tempos distintos conforme a seguir: antes da aplicação dos tratamentos (chamado aqui pré-tratamento), imediatamente após, 24, 48 e 72 horas depois. Cada leitura foi composta pela aquisição de imagens capturadas sequencialmente em intervalos de 250-300 ms – totalizando-se 100 imagens individuais por poço analisado – por meio da aplicação de um microscópio de campo claro automatizado com uma lente objetiva 2x (modelo: *ImageXpress Micro XLS Molecular Devices, CA*). Para testar o efeito de diferentes compostos na viabilidade dos vermes, foram criados grupos de N = 6 poços, com cada grupo recebendo um tratamento diferente, dos quais um foi o grupo controle negativo que foi tratado com o solvente apenas (1-2% DMSO). Com esta configuração experimental, os metadados são: a identificação do poço, o sexo do verme contido no poço, o tratamento/composto, a concentração do composto e o tempo de tratamento.

A análise quantitativa das imagens teve como base a comparação de pares de imagens sequenciais (ao longo do tempo) e foi implementada com o uso do *software open-source CellProfiler*, desenvolvido pelo *BROAD Institute* (CARPENTER et al., 2006; LAMPRECHT et al., 2007), que possui diversos módulos de importação e análise de imagens. Com esta ferramenta, utilizou-se um *pipeline* de análise customizado (construído em trabalhos anteriores do grupo; DANTAS, R. F. et al; *script* em Anexo 1), contendo etapas que aplicam módulos de análise designados para a identificação de poço e correção da luminosidade, detecção de objetos com possibilidade de mascaramento de pixels do plano de fundo da imagem (*background*), entre outras funções (Figura 1).

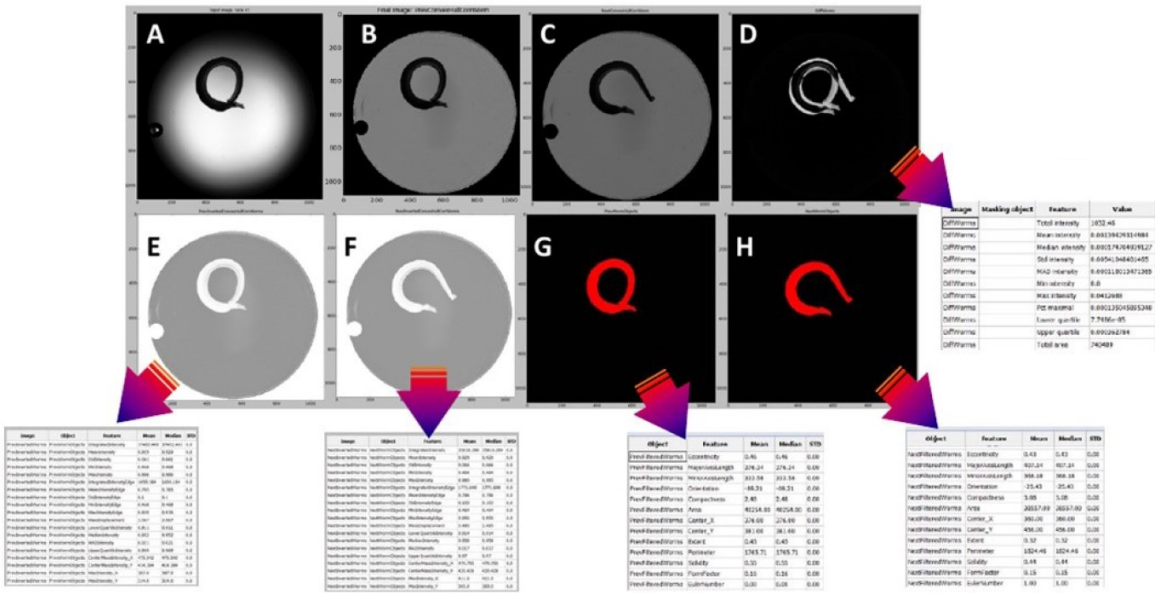


Figura 1 : Algumas etapas do pipeline de análise com vermes adultos de *S. mansoni*. A - Imagem original, B e C – Imagens de um mesmo parasito obtidas num intervalo de 300 ms, com aplicação de filtros para o reconhecimento do poço, remoção de bolhas e correção de iluminação; D - Imagem resultante da diferença absoluta, pixel a pixel, entre as imagens B e C; E e F – Imagens invertidas em intensidade para segmentação e identificação de objetos; G e H – exemplo de par de imagens com *background* mascarado usadas como *input* em módulos de análise.

O processamento das imagens gerou um total de 92 variáveis geradas com o uso do *software CellProfiler,* oriundas de 4 diferentes módulos de medida: *MeasureGranularity*, *MeasureTexture*, *CalculateImageOverlap* e *MeasureImageIntensity*. Os dois primeiros módulos podem conceber medidas relacionadas à textura das imagens, que podem nos dar informações sobre a aparência dos vermes, enquanto que os dois módulos restantes realizam medidas associadas ao tamanho, à opacidade e à motilidade dos parasitas estudados. O *pipeline* para processamento das imagens gerou todas as medidas disponíveis a partir destes 4 módulos de medida.

3.2 Descrição dos dados e do seu processamento

Os dados analisados neste artigo foram coletados no período de 2014 a 2016 pelo nosso grupo de pesquisa, durante o desenvolvimento de uma plataforma de *screening* de alto conteúdo por microscopia em busca de avaliar de forma sistemática, por meio da medição de características relacionadas à motilidade e à

morfologia dos vermes, o efeito de compostos em função do tempo, dose e gênero em parasitas individuais (NEVES et al., 2016a; NEVES et al., 2016b; MELO-FILHO et al., 2016). Neste período acumulou-se dados longitudinais úteis de 1562 vermes machos e 990 vermes fêmeas, antes e depois do tratamento com diferentes compostos.

Por inspeção da descrição dos tipos de medida disponíveis, foi possível obter candidatos para assumir a função de variável-resposta representativa da motilidade a partir dos módulos *CalculateImageOverlap* e *MeasureImageIntensity*, para o desenvolvimento da análise univariada desejada. Foram 4 tipos de medida selecionadas, aplicadas a pares de imagens consecutivas: três, oriundas do módulo *CalculateImageOverlap* e uma obtida com o uso do módulo *MeasureImageIntensity*. Além destas, foi obtida a partir deste último módulo um tipo de medida auxiliar, para cada imagem isolada, chamada *TotalArea*, que pode ser usada como uma medida aproximada do tamanho/área do verme no contexto deste trabalho. A Tabela 1 resume as informações sobre estas variáveis.

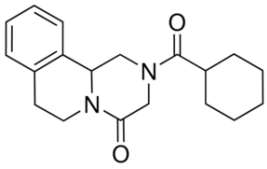
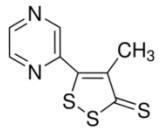
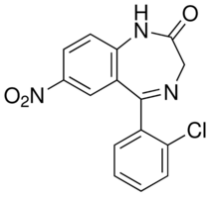
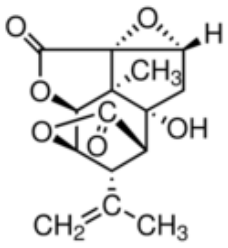
Tabela 1: Descrição das variáveis estudadas na análise da motilidade.

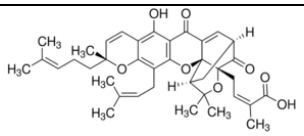
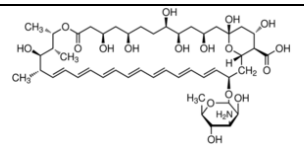
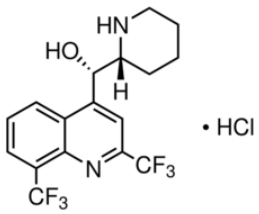
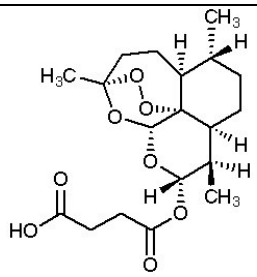
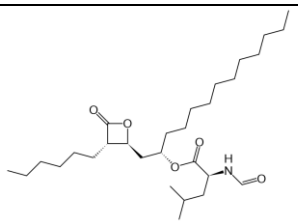
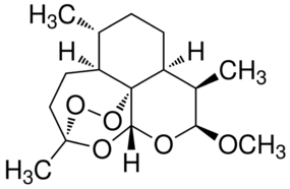
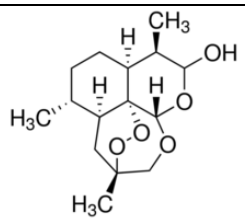
rótulo	Tipo de medida	Módulo	Tipo	Conceituação resumida
FNR	*False Negative Rate*	*CalculateImage Overlap*	Par de imagens com objetos identificados	Fração do número de *pixels* não-mascarados de uma imagem classificados como falso-negativos.
RI	*Rand Index*	*CalculateImage Overlap*	Par de imagens com objetos identificados	Uma medida clássica de similaridade entre agrupamentos, aplicada aos *pixels* de cada imagem, onde cada *pixel* é classificado *foreground* ou *background*
ARI	*Adjusted Rand Index*	*CalculateImage Overlap*	Par de imagens com objetos identificados	É uma versão ajustada da RI, que busca corrigi-la descontando o grau de similaridade que seria esperado ocorrer por acaso.

TI	*TotalIntensity*	*MeasureImage Intensity*	Par de imagens	Somatório, *pixel* a *pixel*, da diferença absoluta entre as intensidades de *pixels* resultante da sobreposição das duas imagens comparadas.
Área	*TotalArea*	*MeasureImage Intensity*	Imagem isolada com objetos identificados	Número total de *pixels* não-mascarados da imagem.

A análise da motilidade desenvolvida neste trabalho foi feita utilizando-se informação da estrutura química dos compostos aplicados aos grupos de tratamento em experimentos de triagem. Os compostos químicos utilizados nos ensaios cujos dados foram a base para a avaliação da modelagem da motilidade estão listados na tabela 2.

Tabela 2: Compostos químicos avaliados na modelagem da motilidade.

Composto	Estrutura	Ação / uso	Refs.	EC_{50}
Praziquantel		Provavelmente metabolismo de cálcio	Ramirez et al., 2007	0,54 - 0,75 µM (120 h)
Oltipraz		Desequilíbrio Redox (glutatione S - transferase)	Nare et al., 1991; Ramirez et al., 2007	ND Ativo em esquistossô-mulos
Clonazepam		Liga-se ao sítio benzodiazepínico do receptor GABA	Noël et al., 2007	ND
Picrotoxina		Antagonista não competitivo do receptor $GABA_A$	Mendonça-Silva et al., 2004	ND 100 µM de picrotoxina produz um fenótipo enrolado e diminui a motilidade

Ácido Gambógico		Agente hemostático Utilizado na medicina chinesa	Peak et al., 2010	ND Ativo em esquistossômulos
Anfotericina B		Droga antifúngica, age no metabolismo do ergosterol	Peak et al., 2010	ND Ativo em esquistossômulos
Mefloquina		Mecanismo exato não é conhecido, forma complexos tóxicos com heme livre gerando dano em membranas	Ingram et al., 2012	11,4 µM
Artesunato		Mecanismo exato não é conhecido, provavelmente age formando radicais livres e outros metabólitos ativos que danificam membranas	Mitsui et al., 2009	ND
Orlistat		Fármaco usando para redução de peso	Panic et al, 2015	ND
Artemeter		Envolve uma interacão com grupo heme ou ions ferrosos, que resultam na produção de espécies reativas	El-Lakkany e Seif El-Din, 2013	ND
Dihidro artemisinina		Mecanismo exato não é conhecido, provavelmente age formando radicais livres e outros metabólitos ativos que danificam membranas	Li et al., 2011	ND

A primeira etapa após a obtenção das medidas foi a limpeza e transformação dos dados colecionados. Realizou-se a padronização da especificação dos metadados, a eliminação de medidas inutilizadas por artefatos experimentais

inerentes ao estudo (como, por exemplo, a não identificação de objetos em uma imagem), além de algumas transformações das variáveis estudadas, necessárias para as análises específicas deste trabalho (transformações logarítmica, *logit* e raiz cúbica). Os *scripts* foram desenvolvidos no *software* livre e ambiente de programação chamado *R* (R CORE TEAM, 2016) e foram escritos tendo em vista a sua aplicação em dados futuros da plataforma de triagem de fármacos. Esta afirmação é válida também para as etapas subsequentes deste trabalho.

3.3 Desenvolvimento, comparação e diagnóstico de modelos estatísticos univariados

Inicialmente, foram realizadas simulações para avaliar o relacionamento básico entre a motilidade e os tipos de medida estudados. Usou-se para isso o modelo mais simples para representar fotos subsequentes de um verme em movimento: representou-se a imagem de dois vermes de tamanhos diferentes como retângulos de dimensões arbitrárias (250 pixels de comprimento e 20 pixels de largura e o outro de 100 pixels por 10 pixels, ambos contidos em poços circulares de diâmetro de 300 pixels). As dimensões foram escolhidas para serem semelhantes às observadas nas bioimagens da figura 1. Para simular o movimento, gerou-se representações deste retângulo após deformação por flexão, como ilustrado na figura 2. Esta deformação é equivalente a um movimento de contração que se observa comumente no comportamento do parasita de *S. mansoni*. Nesta simulação, uma medida de movimento que surge naturalmente é o deslocamento das extremidades dos vermes. Para poder comparar vermes de tamanho diferentes, pode-se definir um *deslocamento relativo* (ao comprimento do verme) como a medida de motilidade de referência (vide figura 2).

As medidas RI e ARI foram obtidas com o auxílio do *R-package* flexclust (LEISCH, 2006). As medidas FNR e TI foram obtidas por contagem de "*pixels"*, levando-se em conta a definição destas. No cálculo da medida TI, tomou-se por uniforme (igual a 1) a distribuição de intensidade do verme simulado e também uniforme (igual a zero) para os pixels de *background* no cálculo da medida TI.

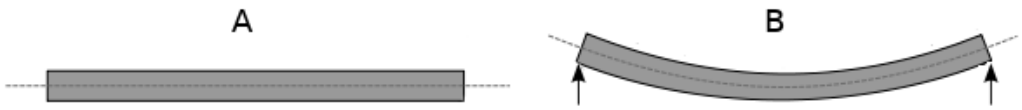


Figura 2: Ilustração dos objetos geométricos utilizados para realizar as simulações de medidas para um movimento simples de deformação por flexão. A - Retângulo original; B - Retângulo deformado. As setas ilustram a extensão do deslocamento das extremidades ao longo do movimento. Se o comprimento do retângulo da figura é unitário, a extensão das setas é o deslocamento relativo.

Em seguida, foi empregada, para cada sexo, uma análise exploratória das quatro variáveis candidatas em representar a motilidade, sujeitas ou não a algumas transformações monótonas, com o intuito de caracterizar a distribuição empírica de cada variável. Foram usadas somente medidas pré-tratamento e de vermes do grupo de controle, para sondar as heterogeneidades que afetam a medida de motilidade na ausência de um composto ativo. Com estes resultados, investigou-se qual variável seria mais adequada como uma medida da motilidade. Em seguida, foram propostos modelos estatísticos da classe dos modelos lineares generalizados multinível (GELMAN e HILL, 2007), compatíveis com as características apresentadas pela variável em estudo. Foram testadas distribuições para variáveis do tipo contagem (beta-binomial, binomial negativa – além das suas versões infladas por zero – e poisson generalizada). A função-*link* empregada para a distribuição beta-binomial (e para sua versão inflada) foi a função *logit*. Tanto para a distribuição poisson generalizada, quanto para a binomial negativa (e para a sua versão inflada por zeros), a função-*link* log foi usada para a ligação entre média μ e preditor linear η, segundo a relação a seguir

$$log\,(\mu) \;=\; log\,(E) \;+\; \eta$$

onde *E* é feito igual à variável *Área*, sendo então log(*E*) o *offset* do modelo, tal que a medida de motilidade média seja proporcional ao tamanho do verme.

A inferência estatística foi feita usando um método INLA - do inglês, *Integrated Nested Laplace Approximation* - (RUE *et al*., 2009), por meio do *R-package* homônimo INLA (LINDGREN *et al*, 2011; MARTINS *et al*, 2013; LINDGREN e RUE, 2015; RUE *et al*, 2017). Para efetuá-la, desta vez utilizou-se dados oriundos de experimentos contendo diferentes compostos em doses fixas ou um único composto testado em diferentes doses, avaliando-se respectivamente modelos para

experimentos de triagem de compostos e ajuste de curvas Dose-Resposta. Para os hiperparâmetros dos modelos, utilizou-se distribuições *a priori* chamadas *PC-priors*[6] (do inglês, *Penalised Complexity* p*riors;* SIMPSON *et al*, 2014). Para os hiperparâmetros foram usadas as distribuições *a priori* chamadas "*pc.prec*", *"pc.gamma"* ou *"pc.mgamma"*, dependendo da identidade do hiperparâmetro (SIMPSON *et al*, 2014). No papel do parâmetro de flexibilidade λ requerido na especificação das distribuições *PC-priors*, usou-se a expressão "λ= - log(0,01)/u", onde "u" é o triplo do desvio padrão estimado da motilidade ao nível de verme na escala *logit*, obtido pelo ajuste de um modelo linear de efeitos mistos aos dados pré-tratamento, com o uso do *R-package* lme4 (BATES *et al*, 2015). Este modelo linear foi ajustado considerando a existência de 3 níveis hierárquicos: medida, verme e experimento (e assim, a especificação de 3 efeitos aleatórios). Com este procedimento, estimou-se para os vermes fêmeas, u = 3,42, e para os vermes machos, u = 3,66.

Na modelagem dos efeitos de tratamento em experimentos do tipo triagem de compostos, foram propostas e testadas duas alternativas para incluir informação química dos compostos: a primeira usando um *Scaffold Network* e a segunda usando-se uma rede de similaridade entre os compostos. Uma terceira modelagem foi utilizada sem inclusão de informação química, para comparação. Como o estudo é longitudinal, foi incluso ao efeito de tratamento um efeito temporal do tipo *random walk* gaussiano de ordem 2, onde o hiperparâmetro variância foi compartilhado entre as replicatas do modelo para cada composto.

Na primeira abordagem (chamada aqui de modelo *Scaffolds*), utilizou-se o *software* gratuito chamado *Scaffold Network Generator* (SNG), que é uma aplicação em linha de comando para a geração de uma rede de compostos, seus *scaffolds* e *subscaffolds* a partir da estrutura química dos compostos estudados (MATLOCK, 2013). Gerou-se *script* em R para utilizar seus comandos automaticamente a partir da interface R, fornecendo como *input* a informação da estrutura química dos compostos codificada na forma de SMILES (do inglês, *Simplified Molecular Input Line Entry Specification*), que é uma forma de representar estruturas químicas

[6] *PC-priors* são definidas como distribuições *a priori* fracamente informativas que se destinam a penalizar a complexidade induzida pelo desvio de um modelo-base mais simples, em busca de combater a tendência ao sobreajuste (do inglês, *overfitting*). Um parâmetro de escala λ (fornecido pelo usuário) é utilizado para controlar a quantidade de flexibilidade permitida pelo modelo (SIMPSON *et al*, 2014).

usando caracteres ASCII. Gerada a rede, sua informação foi incluída no modelo por meio do uso de um modelo BYM (BESAG, YORK E MOLLIÉ, 1991) aplicado à modelagem do efeito de *scaffold.* Nesta abordagem, o efeito de composto é definido como a soma de um efeito de *scaffold* (compartilhado entre os compostos que possuem um mesmo *scaffold*) e de um efeito extra para cada composto. Usou-se, por simplicidade, o modelo IID gaussiano (centrado em zero e variância desconhecida) para modelagem destes efeitos extras. Neste trabalho, as relações de vizinhança "espacial" entre os *scaffolds* foram definidas em termo dos *sub-scaffolds* primários: dois *scaffolds* são ditos *vizinhos* quando eles compartilham um mesmo *sub-scaffold* primário a ambos os *scaffolds*. Desta forma, o efeito do tratamento na motilidade é a soma de 3 efeitos aleatórios:

$$\beta_{(c,t)} \;=\; \beta_c^{(un)} \;+\; \psi_{s[c]}^{BYM} \;+\; \gamma_{(t\times c)}^{RW2}$$

$$\beta_c^{(un)} \;\sim\; N\left(0\,,\; \sigma^2_{\beta^{(un)}}\right)$$

onde é $\beta_{(c,t)}$ o efeito de tratamento do composto *c* após o tempo *t;* $\beta_c^{(un)}$ *é o* efeito extra de cada composto; $\gamma_{(t\times c)}$ é o efeito do tempo RW2 para cada composto separadamente; e $\psi_{s[c]}$ é o efeito do *scaffold s* referente ao composto *c*. Este último é especificado como a soma de um efeito IID e outro efeito estruturado $\psi^{(st)}_{s[c]}$, cuja distribuição condicional é gaussiana, com parâmetros média e variância na forma

$$\psi_j^{(st)} \,|\, \psi_{k\neq j}^{(st)} \;\sim\; N\left(\tfrac{1}{n_j}\textstyle\sum_{j\sim k}\psi_k^{(st)}\;,\; \tfrac{1}{n_j}\sigma^2_{\psi^{(st)}}\right)$$

onde n_j é o número de *scaffolds* vizinhos ao *scaffold j* e " $j \sim k$ " indica que dois *scaffolds j* e *k* são vizinhos.

Na segunda abordagem (nomeada modelo Tanimoto), utilizou-se o *R-package ChemmineR* (HORAN, 2018), que é um pacote de ferramentas de quimioinformática para aplicações em ambiente R. Com auxílio deste pacote, foram obtidos coeficientes de similaridade Tanimoto, com base em descritores AP (*atom pair, em inglês;* CARHART et al, 1985; CHEN e REYNOLDS, 2002). Um modelo BYM ponderado foi usado na modelagem do efeito de composto, empregando-se para isso a rede de compostos ponderada com base nas similaridades calculadas. Desta forma, a dependência "espacial" no modelo BYM foi especificada pela matriz de coeficientes de similaridade Tanimoto. Nesta segunda abordagem, o efeito do tratamento é a soma de 2 efeitos aleatórios:

$$\beta_{(c,t)} = \beta_c^{BYM} + \gamma_{(t\times c)}^{RW2}$$

onde a notação é a mesma da anterior.

Distintamente, a modelagem sem inclusão de informação química foi feita usando um modelo IID gaussiano (centrado em zero e variância desconhecida) para o efeito do composto, que é a versão bayesiana para se obter coeficientes de regressão “encolhidos” em torno de zero como solução do método *ridge regression* (JAMES et al, 2013). Ela é referida daqui pra frente no texto como “modelo básico”.

Estes modelos foram aplicados aos dados de um experimento que possuem informação sobre a estrutura química dos compostos testados como um metadado extra, codificado na forma de SMILES. Os compostos sob análise, dentre os quais incluem fármacos utilizados para outras doenças, foram testados em concentração fixa de 50 μM; no total, 10 compostos foram testados durante a realização deste experimento: ácido gambógico, Anfotericina B, Artemeter, Artesunato, Clonazepam, Dihidro-artemisinina, Mefloquina, Orlistat, Picrotoxina e Praziquantel (o controle-positivo). Este experimento foi composto por 47.338 medidas de motilidade, das quais 18.240 são de vermes fêmeas e 29098 de vermes machos distribuídos entre os 11 grupos de tratamento (incluindo o grupo controle). As medidas provêm da observação de 57 parasitos fêmeas e de 69 parasitos machos.

Na modelagem da relação Dose-Resposta, foram empregados modelos com o intuito de se construir curvas dose-resposta para a motilidade média de dois tipos comumente estudados (RITZ et al, 2015): uma curva dose-resposta *Weibull I* de quatro parâmetros (que permite curvas assimétricas) e uma curva log-logística generalizada de quatro parâmetros (simétrica)[7].

Para ajustar a curva *Weibull I,* o efeito do composto (em função da dose) foi construído como um modelo latente *Ornstein-Uhlenbeck,* tomando-se fixo o hiperparâmetro precisão marginal τ = exp(28). A ele, um hiperparâmetro de escala desconhecido adicional (restrito a ser positivo) foi incluído com o uso da chamada *copy-feature* (MARTINS et al, 2013), tomando-se x_1 = -1 e definindo-se a variável preditora como o cologaritmo de *D*, onde *D* é a razão entre a dose e a maior dose

[7] Os quatro parâmetros de ambos os modelos são: um parâmetro “*b*”, que é o parâmetro de *slope* da curva; um parâmetro “c”, que corresponde ao limite inferior de motilidade; um parâmetro “*d*”, que é o limite superior de motilidade (e que define a motilidade do grupo controle); e um parâmetro “*e”* (ou EC_{50}) é a dose associada à redução da motilidade para metade da diferença d - c. Os modelos descritos a seguir foram construídos para o caso particular onde c = 0.

em estudo. Isto foi assim implementado para mudar a unidade de concentração da dose tal que variável preditora seja zero quando a dose for máxima, e o modelo resultante para o efeito de composto seja

$$\beta_t = -\left(\frac{dose}{I_t}\right)^{b_t} + \gamma_t^{RW2}$$

onde o efeito do tempo permanece igual ao existente no modelo para um experimento de triagem, rejeitando-se o índice *c*, já que há somente um composto estudado. Os parâmetros *b* e *I* são avaliados para cada tempo estudado (o hiperparâmetro de escala adicionado estima $(I)^{-b}$ enquanto *b* é estimado diretamente; a distribuição simulada de *I* é obtida a partir da distribuição conjunta de *b* e do hiperpârametro de escala). Usando-se a função-*link* log, esta especificação para o efeito do tratamento é equivalente a assumir uma relação dose-resposta do tipo *Weibull I* para a média da resposta de motilidade, onde os parâmetros *b* e *I* são, respectivamente, os parâmetros de forma e o ponto de inflexão da curva (RITZ, 2015). Para cada tempo separadamente, um parâmetro EC_{50} é obtido pela identidade $EC_{50} = I \cdot log(2)^{1/b}$ (RITZ et al, 2015).

O ajuste da curva log-logística generalizada foi possível com o uso do modelo latente *log1exp* (LINDGREN e RUE, 2015). O hiperparâmetro de escala nativo a este modelo foi mantido fixo e igual a -1, enquanto a variável preditora foi tomada como o cologaritmo da dose. Como resultado, o efeito de composto, em função do tempo e da dose, é dado pela expressão

$$\beta_t = -\log(1 + \exp(b_t(\log(dose) - \log(e_t)))) + \gamma_t^{RW2}$$

onde o parâmetro e_t corresponde à concentração do composto que induz metade do efeito máximo, comumente abreviado como EC_{50}; enquanto o parâmetro b_t é o parâmetro de *slope* da curva. De fato, os parâmetros diretamente estimados usando-se o modelo latente *log1exp* são o próprio b_t e o parâmetro nativo $\alpha_t = -b_t \log(e_t)$, de onde o e_t pode ser derivado. Esta derivação foi feita, para cada tempo, por simulação da distribuição conjunta de α_t e b_t , resultando em simulações da distribuição de e_t após o emprego da transformação citada acima.

Tanto para a curva Weibull quanto para a curva log-logística generalizada, o parâmetro d_t é obtido a partir da identidade $d_t = \exp(\alpha + \beta_{t,controle})$, onde α é um efeito fixo incluso ao preditor linear e $\beta_{t,controle}$ representa o efeito β_t com dose = 0 (ou seja, o grupo controle).

Gerou-se também *scripts* para testar uma generalização dos modelos dose-resposta anteriores que contemplem o efeito de hormesis[8], do tipo Cedergreen-Ritz-Streibig (CEDERGREEN et al, 2005; RITZ et al, 2015). Uma versão modificada deste modelo foi produzida por adicionar ao efeito de tratamento β_t um termo extra do modelo latente *log1exp* (LINDGREN e RUE, 2015). Isto foi feito tomando-se a variável x igual ao inverso da raiz quadrada da dose, o parâmetro de escala $\theta = 1$ e o parâmetro $\gamma = 0{,}02$.[9]

Estes modelos para dose-resposta foram testados com dados de experimento cujo tratamento foi a adição de um composto químico antiparasitário chamado oltipraz[10], onde a faixa de concentração estudada foi de 0,005 μM até 10 μM (com 8 concentrações). Este experimento contém dados de um total de 100 vermes, dos quais 44 são vermes fêmeas e 56 machos, sendo 15 vermes do grupo controle (6 fêmeas e 9 machos).

Foram testados três modelos para a heterogeneidade e relações de dependência entre os efeitos esperados devido à variação a nível de verme:

i. O primeiro modelo atribui somente um efeito aleatório IID gaussiano por verme, centrado em zero e hiperparâmetro variância desconhecida (chamado aqui: modelo IID).

ii. O segundo modelo testado é constituído por dois termos: o primeiro termo é idêntico ao modelo anterior, enquanto o segundo termo é um efeito IID (gaussiano centrado em zero e hiperparâmetro variância desconhecida), considerado para a interação verme × tempo. Sendo assim, o efeito de verme neste caso é modelado como:

$$\delta_{(w,t)} = \delta_w^{(0)} + \varepsilon_{w\times t}$$

onde o somatório dos efeitos $\varepsilon_{w\times t}$ de um mesmo verme w é restrita a zero, para cada verme incluído no modelo. Esta restrição é feita para que os efeitos $\varepsilon_{w\times t}$ sejam análogos a um termo de erro em relação ao efeitos de verme $\delta^{(0)}$, além de evitar

[8] Hormesis caracteriza um processo bifásico: há um aumento da resposta com o aumento da dose, para doses baixas (fase 1); e um processo de tendência inversa, para doses maiores (fase 2).

[9] A raiz quadrada (ou seja, expoente igual a 0,5) e o parâmetro $\gamma = 0{,}02$ foram arbitrariamente definidos por uma avaliação prévia dos dados em estudo. Esta escolha prévia de parâmetro é uma característica deste modelo Cedergreen-Ritz-Streibig. A única modificação feita sobre o modelo original foi a definição do parâmetro $\gamma \neq 1$. Estes parâmetros juntos governam a taxa de aumento do efeito hormesis e eles são pré-definidos somente para não superparametrizar o modelo final (CEDERGREEN et al, 2005).

[10] Nome IUPAC: 4-metil-5-(2-pirazinil)-3-ditioletiona.

problemas com não-identificabilidade dos efeitos/parâmetros. Este segundo modelo é referido aqui como modelo IID+IID.

iii. O terceiro modelo foi construído atribuindo-se ao efeito de verme uma estrutura de erro espacial SEM (BIVAND et al, 2014). Seja N o número de vermes, T_w o número de ensaios realizados com o verme w, e m o somatório dos T_w para w = 1, 2, 3, ..., N; então a matriz de pesos W utilizada foi uma matriz esparsa $m \times m$, obtida a partir da padronização por linhas de uma matriz com os elementos da diagonal principal todos nulos, enquanto que elementos fora da diagonal são nulos caso correspondam a vermes diferentes, e igual a um caso contrário. Os hiperparâmetros deste modelo são a variância do ruído gaussiano e um termo de correlação ρ_{Err} (BIVAND et al, 2014), permitido assumir valores no intervalo [0, 1], sendo este último interpretado como a correlação entre os efeitos $\varepsilon_{w \times t}$ para cada verme w, em média. Neste modelo, $\delta^{(0)}$ não foi incluído, o que resulta na identidade $\delta_{(c,t)} = \varepsilon_{w \times t}$ (isto é, δ é puramente o modelo SEM). Este último modelo é referido aqui como modelo SEM.

A comparação entre os modelos testados foi guiada pelo critério de informação de Watanabe-Akaike (WAIC, do inglês *Widely Applicable Information Criterion* ou *Watanabe-Akaike information criterion*), que é uma alternativa rápida e conveniente computacionalmente à técnica de validação cruzada (GELMAN *et al*, 2014). Além disso, foi adotada uma abordagem baseada na extração de resíduos escalados (DUNN e SMYTH, 1996) [11], com o intuito de diagnosticar a adequação dos modelos aos dados. Sua implementação foi feita com o uso do *R-package* chamado DHARMa (HARTIG, 2018), produzindo resíduos escalados a partir de simulações da distribuição *a posteriori* preditiva correspondente a cada modelo ajustado, obtidas com o uso do *package* INLA.

3.4 Desenvolvimento de modelos preditivo multivariado

Com a intenção de construir um modelo multivariado semelhante ao obtido para o caso univariado - e usando-se de métodos já consagrados de análise multivariada - seguiu-se, em resumo, o seguinte procedimento:

i. transformação das variáveis;

[11] Eles são prontamente interpretáveis, com a clareza que a abordagem gráfica permite.

ii. redução de dimensionalidade por PCA;
iii. agrupamento fuzzy dos dados;
iv. concluindo em um modelo preditivo do tipo árvore de regressão (com efeitos aleatórios estimados a nível de verme). Para tal, empregou-se o coeficiente de pertencimento ao *cluster* controle[12] como a variável-resposta.

Inicialmente, um estudo exploratório foi realizado para avaliar o grau de correlação entre as variáveis e formas das suas distribuições, com o intuito de propor transformações úteis na geração de variáveis transformadas com distribuição univariada mais simétrica (a fim de melhorar os resultados de PCA). Nesta etapa, foram utilizados dados de vermes pré-tratamento, para cada sexo separadamente. A transformação *raiz cúbica* foi aplicada às variáveis assimétricas para a direita. As variáveis que apresentam assimetria para a esquerda foram transformadas linearmente em variáveis com assimetria para direita, e após isso submetidas à raiz cúbica[13].

Com os dados de pré-tratamento transformados, realizou-se a análise de componentes principais, extraídos via matriz de correlação, para cada sexo. Isto foi feito com o uso da função *princomp* do *R-package stats.* Observações incompletas (isto é, com um ou mais valores ausentes) foram retiradas do conjunto de dados.[14] O número de componentes principais retidos foi decidido usando o critério de *Kaiser* (isto é, foram mantidos somente componentes principais com autovalores maior que a unidade). A partir do subconjunto selecionado dos componentes principais obtidos, foram derivadas as funções-*score* para enfim reduzir a dimensionalidade dos dados de tratamento. Os dados com dimensão reduzida são usados nas etapas subsequentes.

O próximo passo foi um estudo sistemático da estabilidade de resultados de agrupamento *fuzzy*, em função dos parâmetros de entrada do método, para cada tempo e sexo separadamente. A abordagem baseada em estabilidade aqui empregada é um critério objetivo para comparar diferentes soluções de

[12] O *cluster* controle é definido aqui neste trabalho como aquele ao qual *pertence* o maior número de observações advindas de vermes do grupo controle negativo experimental.

[13] Em símbolos: Se a variável X é assimétrica positiva, a transformação usada foi $\sqrt[3]{X}$; Se a variável é assimétrica negativa, a transformação usada foi $\sqrt[3]{(1 - X)}$.

[14] Os números de observações incompletas foram somente 5 (de um total de 83486 medidas de motilidade) para vermes fêmeas e 11 (de um total de 141773) para os vermes machos.

agrupamento e identificar a mais robusta delas e o número de *clusters* adequado (LANGE et al, 2004)[15].

Para cada sexo, os dados foram divididos em dois grupos, chamados grupo de treinamento e o outro, grupo de teste. Cada grupo recebeu dados de metade da coleção de vermes de um mesmo sexo (sendo os vermes de cada grupo alocados aleatoriamente, gerando dois extratos). A separação dos dados foi feita a nível de verme[16]. A partir de cada um dos dois grupos foram gerados 20 subgrupos disjuntos, agora a nível de medida, onde cada um deles possui uma única observação por verme. Com o uso de cada um destes subgrupos foram gerados resultados de agrupamento *Fuzzy C-Means* - usando o algoritmo FANNY (KAUFMAN e ROUSSEEUW, 2009), por meio do *R-package cluster* (MAECHLER et al, 2018) - separadamente para cada sexo e tempo após tratamento (0h, 24h, 48h, 72h), e atribuindo-se diferentes valores para os parâmetros de entrada: *membership exponent r* (r = 1,1; 1,125; 1,15 e 1,175) e o número k de *clusters* (de 2 a 6). A medida de distância adotada foi a distância de *Mahalanobis*, com o intuito de permitir a detecção de *clusters* de formato elipsoidal (MAHALANOBIS, 1936). Ela foi calculada com auxílio da função *fDiss* implementada no R-package *resemble* (RAMIREZ-LOPEZ e STEVENS, 2016).

A validação baseada em estabilidade é feita por meio das etapas a seguir:

a) construção de um modelo preditivo (ou uma função preditiva), usando os resultados de agrupamento obtidos a partir dos dados de treinamento;

b) aplicação deste aos dados de teste;

c) e finalmente a comparação entre os valores preditos e aqueles obtidos por análise de agrupamento aos dados de teste (para mais detalhes, vide LANGE et al, 2004).

Neste trabalho, construiu-se uma função preditiva baseada na função objetivo do algoritmo FANNY (KAUFMAN e ROUSSEEUW, 2009; MAECHLER et al, 2018).

---

[15] A abordagem baseada em estabilidade é essencialmente uma extensão do método de validação-cruzada aplicável à análise de agrupamento. O critério de estabilidade do agrupamento explora a ideia de que quando múltiplos subconjuntos de dados são amostrados a partir de uma mesma distribuição, espera-se que o algoritmo de agrupamento empregado se comporte de forma semelhante e produza resultados similares entre as amostras.

[16] Isto foi feito assim porque é mais relevante avaliar a estabilidade do agrupamento quando se adiciona um novo verme ao conjunto de dados, ao invés de uma nova observação de um verme já considerado no agrupamento.

Para isso, a função objetivo original é resgatada a partir dos coeficientes de pertencimento estimados e reminimizada para uma nova observação incluída, mantendo-se fixos os coeficientes de pertencimento dos dados de treinamento. Este procedimento foi implementado em ambiente R (para ver o *script* usado, vide Anexo 2), onde a minimização foi feita com o uso da função *constrOptim.nl* do *R-package alabama* (VARADHAN, 2015). Para cada configuração estudada, foram obtidas medidas do erro quadrático médio (RMSE, do inglês, *Root-Mean-Square Error*) entre os valores estimados por análise de agrupamento e valores preditos, das 20 replicatas tomadas.

Após a escolha da configuração de parâmetros para a análise de agrupamento, foram obtidas estimativas dos coeficientes de pertencimento para todas as observações dos dados após tratamento, por sexo e tempo de tratamento. O *cluster* controle foi identificado[17] e os coeficientes de pertencimento a ele foram usados como variável-resposta na construção de um modelo do tipo árvore de regressão (um para cada tempo e sexo). Estes modelos incluem efeitos aleátorios e correlação entre as replicatas de um mesmo verme, graças a combinação entre um algoritmo de estimação de árvore de regressão simples e outro algoritmo para ajuste de modelo de regressão linear com efeitos aleatórios (SELA e SIMONOFF, 2012). Isto foi feito com o uso do *R-package REEMtree* (SELA e SIMONOFF, 2015). Os modelos foram ajustados utilizando-se os resultados do grupo de treinamento e do grupo de teste separadamente. Diferentemente da etapa de agrupamento *fuzzy*, nesta última etapa os dados *de treinamento* foram compostos pela primeira metade de observações de cada verme, enquanto os dados *de teste* corresponderam à outra metade restante. Em seguida, foi feito um estudo de validação cruzada para avaliar a consistência interna[18] na capacidade de previsão dos modelos. Para isso foram calculados RMSE e comparados aos valores estimados do desvio padrão dos resíduos dos modelos.

[17] A identificação do *cluster* controle foi feita por meio da determinação do *cluster* ao qual *pertence* a maioria das medidas do grupo controle negativo.

[18] Como a diferença entre os dados de treinamento e os dados de teste é a nível de medida (e não a nível de verme), esta avaliação busca somente verificar se as observações de cada verme produzem resultados coerentes para cada verme. Análises adicionais, buscando avaliar o poder preditivo do modelo árvore de regressão com efeitos aleatórios para vermes não contidos no grupo de treinamento, não foram realizadas neste trabalho; ela foi deixada para um trabalho futuro que complemente esta dissertação.

4 RESULTADOS E DISCUSSÃO

4.1 Análise Exploratória Univariada

Na primeira parte deste projeto, a investigação esteve centrada na busca de modelagem adequada aos dados associados à motilidade dos parasitas, porque esta característica fenotípica é um indicador bem estabelecido na literatura para a avaliação do efeito de fármacos sobre helmintos (BUCKINGHAM *et al*, 2014). A motilidade pode servir como uma medida indireta da saúde do parasita e da sua atividade neuromuscular. Todavia, a medida de motilidade é resultante de diversos fatores, tanto aqueles inerentes ao comportamento dinâmico do animal sob observação quanto outros relacionados à configuração experimental em si. Estes fatores se conjugam (não necessariamente de forma linear e aditiva) e aumentam a complexidade da variável-resposta; sem, no entanto, nenhuma garantia de que a incerteza a ela associada possua características que a enquadrem em uma ou outra distribuição de probabilidade conhecida.

Apesar disso, infelizmente a preocupação com a validação dos modelos estatísticos empregados é ainda muito tímida no meio acadêmico. Por exemplo, é ainda muito comum assumir erroneamente uma distribuição normal aos erros associados a uma variável-resposta que apresente forte assimetria na distribuição de erros ou clara relação funcional entre variâncias e médias estimadas em diferentes estratos da amostra. Quando isto ocorre, tanto os valores estimados e intervalos de confiança associados não são confiáveis, quanto o poder explicativo da análise efetuada pode ser bastante prejudicado; isto é, o modelo resultante pode ser enganoso não somente em termos quantitativos, mas também qualitativamente (GELMAN e HILL, 2007).

Uma abordagem comum em análise de dados é o uso de dados agregados. Neste trabalho, poder-se-ia agregar as replicatas por indivíduo a cada tempo estudado e utilizar a média das replicatas como representativa da motilidade. Nesta configuração, cada verme de um mesmo grupo de tratamento, em um dado tempo de ensaio após o tratamento, desempenharia a função de replicata para a medida de motilidade do seu grupo. Graças ao Teorema Central do Limite, podemos ter a partir desta abordagem uma maior confiança (apesar de ainda nenhuma garantia) na

aplicabilidade de métodos baseados na suposição de erros gaussianos, como a popular análise de variância (ANOVA, ou uma de suas generalizações), mesmo que a variável original não seja normalmente distribuída. Entretanto, o uso de dados agregados neste trabalho não é a abordagem mais interessante, dada a complexidade do tipo de parasita em estudo. Isto acontece principalmente porque o número de replicatas por grupo é restrito a N = 6, pois existem limitações experimentais e financeiras que restringem o número de vermes disponíveis em cada ensaio. Desse modo, se as fontes de heterogeneidade possuem grande impacto na medida de motilidade média por indivíduo, o poder de discriminação entre grupos em uma análise tipo ANOVA seria muito reduzido pelo pequeno tamanho da amostra. Ao final da seção 4.2.1, é feita uma comparação entre os resultados obtidos usando dados agregados e os resultados com o uso do modelo proposto neste trabalho.

Para isso, neste trabalho é investido tempo na análise exploratória dos dados e na avaliação do emprego de modelos lineares generalizados multinível aos dados experimentais, em busca do desenvolvimento de um modelo estatístico robusto que descreva adequadamente os dados experimentais ao nível de medida/replicata e que possa superar as limitações existentes em uma análise do tipo ANOVA com dados agregados. Além disso, os resultados desta modelagem já permitem diretamente a análise da intensidade do efeito de cada tratamento, sem a necessidade de aplicação de testes *post hoc* (diferentemente da ANOVA).

4.1.1. *Conceituação e simulação das variáveis candidatas a representar a motilidade*

As quatro variáveis estudadas são medidas relacionadas ao grau de atividade neuromuscular que o parasita apresenta em um pequeno intervalo de tempo entre a aquisição de uma fotomicrografia e outra, através de medidas que são funções da semelhança entre as imagens. A variável TI representa a diferença absoluta de intensidade de pixel de uma imagem e a outra (vide a figura 1.D); como característica básica, o valor mínimo possível para TI é zero, indicando que nenhum movimento pôde ser detectado.

As variáveis obtidas pelo módulo *CalculateImageOverlap* (FNR, RI e ARI) são resultado da sobreposição de um par de imagens exemplificadas na figura 1.G-H e a classificação dos pixels contidos nos objetos identificados. A variável FNR é simplesmente a proporção de *pixels* do objeto da primeira imagem que não pertencem ao objeto da segunda imagem, servindo como medida aproximada da fração de segmentos do corpo do animal que apresentou movimento. Por sua vez, no contexto da análise de imagens, as variáveis RI e ARI são auferidas a partir de cálculos mais complexos, que geram índices da similaridade entre duas imagens consecutivas (RAND, 1971; HUMBERT e ARABIE, 1985). Como o interesse é medir a dissimilaridade entre as imagens, é útil obter o complemento para a unidade das variáveis RI e ARI, respectivamente (isto é, CRI = 1 - RI e CARI = 1 - ARI). O complementar do Rand Index (CRI) é a proporção de pares de pixels que satisfazem uma propriedade: um dos pixels do par corresponde a um “ponto” da imagem do parasita que aparentemente manteve-se em repouso, enquanto que o outro pixel corresponde a um outro “ponto” que teve seu movimento detectado. A variável CARI é uma medida que busca corrigir a medida CRI, assumindo-se a distribuição hipergeométrica generalizada (MATHAI e SAXENA, 1967) como modelo de aleatoriedade da classificação dos pixels. Logo, elas também servem, a princípio, como uma medida do grau de atividade aparente do animal no intervalo de tempo entre a aquisição de uma imagem e outra.

Com o intuito de alcançar uma visão mais nítida da relação entre a motilidade de um parasita e os tipos de medida analisados, foram feitas simulações das 4 medidas estudadas, variando-se o deslocamento relativo das extremidades dos vermes simulados (vide figura 2), que é usada na simulação como medida de referência da motilidade. A figura 3 resume os resultados de simulação.

Primeiramente, nota-se que TI é um múltiplo de FNR para os dados simulados, onde a razão α entre eles é dada pelo dobro da área do retângulo de cada verme simulado (10^4 *pixels* para o verme maior e $2 .10^3$ pixels para o menor). Além disso, são somente estas duas variáveis que apresentam uma relação exatamente linear com a variável deslocamento relativo, dentro de uma faixa central de valores. Por sua vez, as variáveis CRI e CARI são só aproximadamente lineares em relação à variável deslocamento, sendo a primeira a mais problemática, pois só assume valores muito baixos (o que pode implicar negativamente na sensibilidade do método experimental).

Já era esperado que a variável CRI sofresse forte influência do tamanho dos objetos enquanto CARI fosse menos sensível a isso (HUMBERT e ARABIE, 1985). De fato, o ajuste feito em CRI para gerar a variável CARI tem o efeito benéfico de reescalá-la e diminuir o efeito do tamanho dos objetos sobre ela; entretanto, a comparação dos resultados simulados de vermes/objetos de tamanhos diferentes (ilustrada na figura 3) demonstra que o ajuste por chance que a define não foi suficiente para tornar a medida CARI invariante ao tamanho do objeto. Ao contrário, FNR e TI (reescalado) são de fato invariantes ao tamanho do objeto e equivalentes entre si, para dados obtidos por simulação.

Como um problema geral, todas as variáveis desviam fortemente da linearidade em relação ao deslocamento relativo quanto tendem a zero ou ao valor máximo de cada variável. O limite superior é uma limitação do próprio método de medida, que envolve a sobreposição de objetos presentes nas imagens; isto é, quando ocorrem deslocamentos muito grandes, a sobreposição pode ser nula (e assim, independente do valor real do deslocamento), ocasionando o surgimento de dados censurados à direita. Na prática, isto não deve ser um problema, pois a configuração experimental foi planejada com intuito de mensurar pequenos deslocamentos. O problema da não-linearidade próximo a zero é mais importante, porque a faixa de interesse na análise da motilidade inclui os valores nulos, muito importantes para sondar a inatividade ou morte do parasita. Investigando a origem deste problema, realizou-se simulações idênticas às aqui apresentadas, diferindo somente na largura dos retângulos simulados, que foi tomada igual a unidade (por simplicidade, os resultados não foram ilustrados). Neste modelo de verme “linear”, a relação entre deslocamento, FNR e TI é exatamente linear para toda a faixa de valores, mostrando que o problema observado é consequência da largura não desprezível dos vermes, quando comparada à extensão do deslocamento. Logo, existe um limite de quantificação do deslocamento relativo quando forem usadas qualquer um dos tipos de medida em estudo.

Nesta análise inicial, resulta que as opções de medida com características mais desejáveis para representar a motilidade são as variáveis FNR ou TI, ou transformações destas.

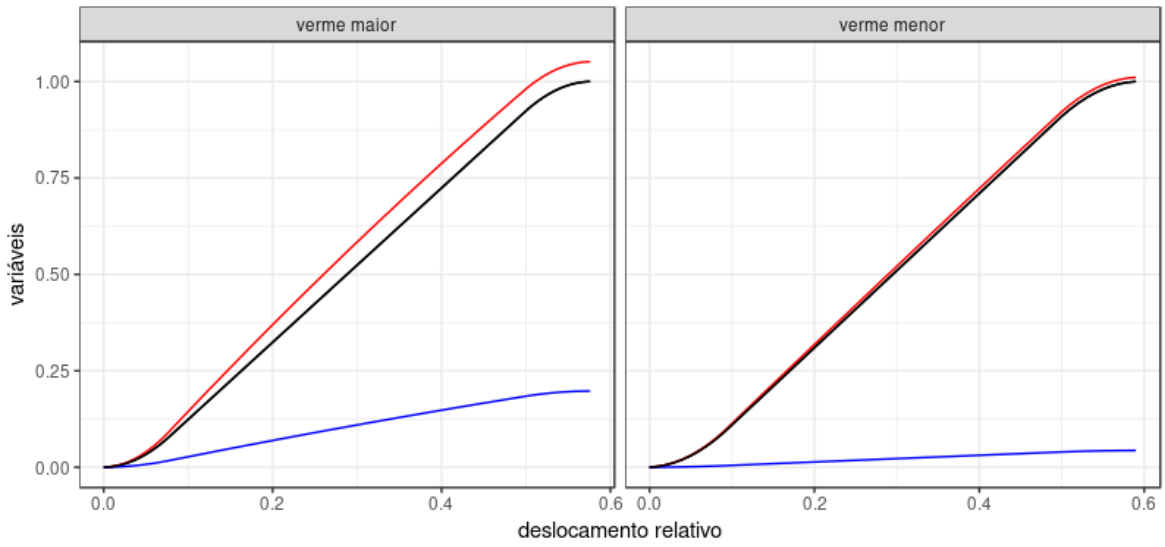


Figura 3: Relação entre as variáveis simuladas e o deslocamento das extremidades de um verme simulado relativo ao seu comprimento. Preto - curvas sobrepostas das variáveis FNR e TI reescalada (TI reescalada = TI / α, onde α é o dobro da área do verme); azul - variável CRI, vermelho - variável CARI.

4.1.2. *Análise exploratória dos dados pré-tratamento e de vermes do grupo controle.*

Antes de buscar um modelo que relacione a motilidade dos parasitas e o efeito do tratamento, é preciso identificar quais as características que as variáveis relacionadas à motilidade apresentam na ausência de tratamento em dados reais, e assim identificar como outros efeitos podem influenciar a análise da motilidade com o uso de cada tipo de medida. É natural começar esta análise pela avaliação dos histogramas das variáveis candidatas. Na figura 4, temos o perfil de distribuição da variável FNR para dados pré-tratamento, onde se pode notar uma forte assimetria. Ademais, é evidente (e em especial para vermes machos) que a frequência observada cai exponencialmente com o aumento do valor da variável próximo ao seu valor mínimo, zero, até o valor máximo igual a 1. Claramente, este perfil é muito diverso do que seria observado caso esta variável fosse distribuída normalmente ou por uma mistura de gaussianas.

Como esperado pelos resultados de simulação, a distribuição das variáveis restantes tem perfil muito semelhante à distribuição da variável FNR, com exceção da escala (vide figuras 5-7). Nesta amostra estudada de dados pré-tratamento, a variável FNR foi a única a apresentar valores exatamente nulos e valor máximo igual

a 1; a variável CRI foi limitada a valores menores que 0,2 enquanto que CARI apresentou valor máximo de 1,04, ambas assumindo valores estritamente positivos.

A figura 7 ilustra que a variável TI apresenta valores extremos muito mais acentuados que as demais variáveis. Isto provavelmente ocorre em função de artefatos presentes em algumas imagens que não afetam as medidas obtidas com o uso do módulo *CalculateImageOverlap*, mas que podem afetar a variável TI porque esta última não é auferida com o uso de imagens com objeto identificado (CARPENTER *et al.*, 2006; LAMPRECHT *et al.*, 2007).

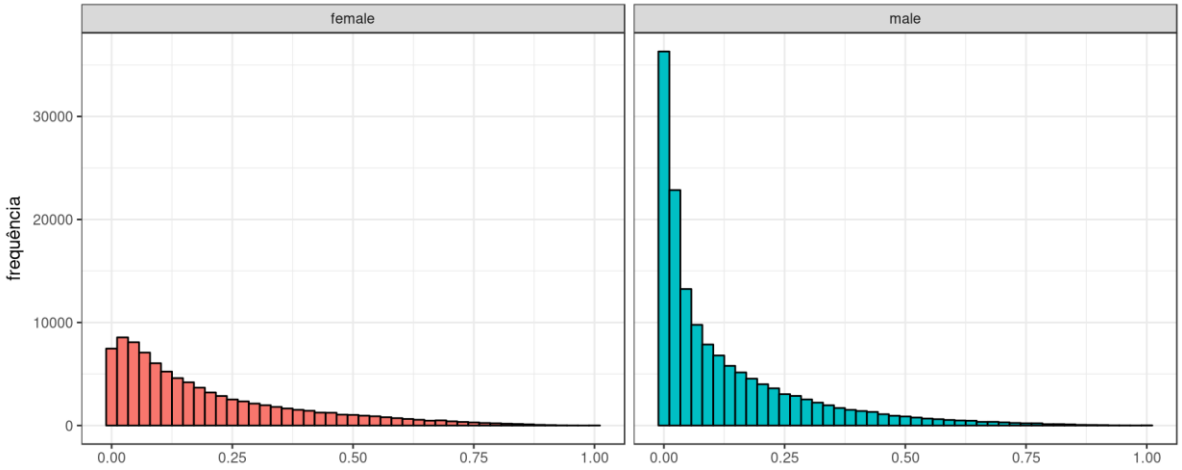


Figura 4: Histogramas da variável FNR para medidas pré-tratamento de parasitas *S. mansoni*. Amostra: 88.527 medidas de fêmeas e 147.228 medidas de machos.

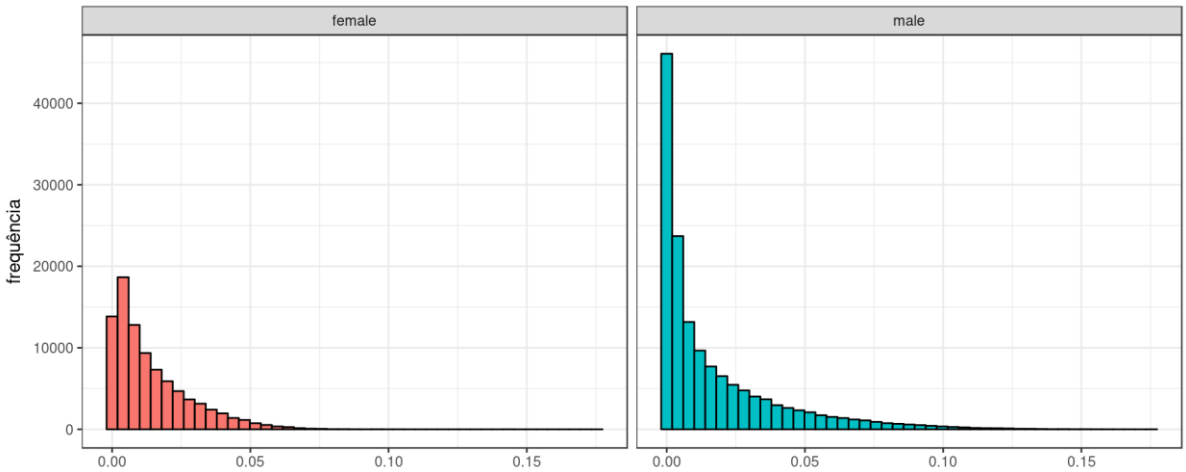


Figura 5: Histogramas da variável CRI para medidas pré-tratamento de parasitas *S. mansoni*. Amostra: 88.527 medidas de fêmeas e 147.228 medidas de machos.

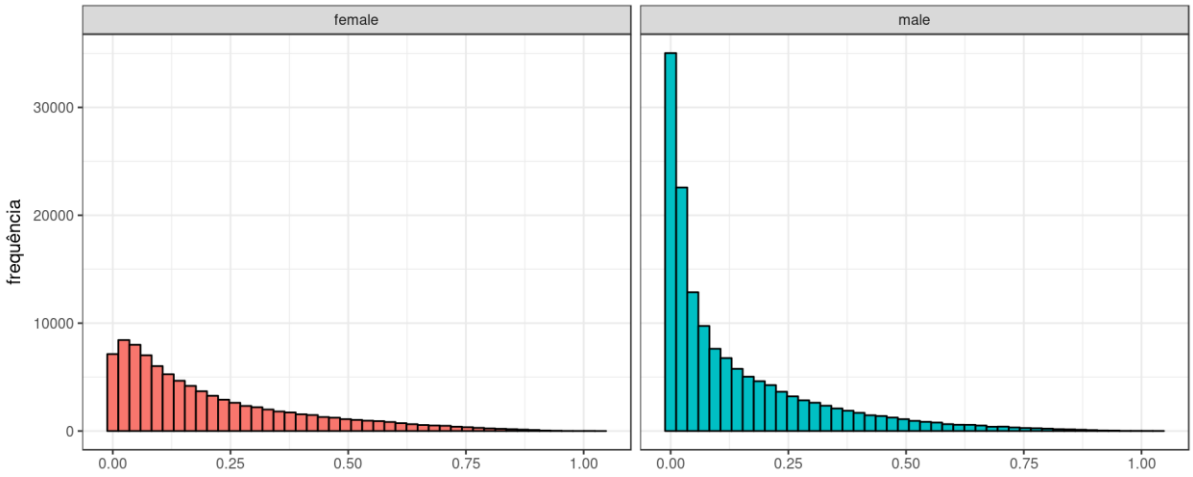


Figura 6: Histogramas da variável CARI para medidas pré-tratamento de parasitas *S. mansoni*. Amostra: 88.527 medidas de fêmeas e 147.228 medidas de machos.

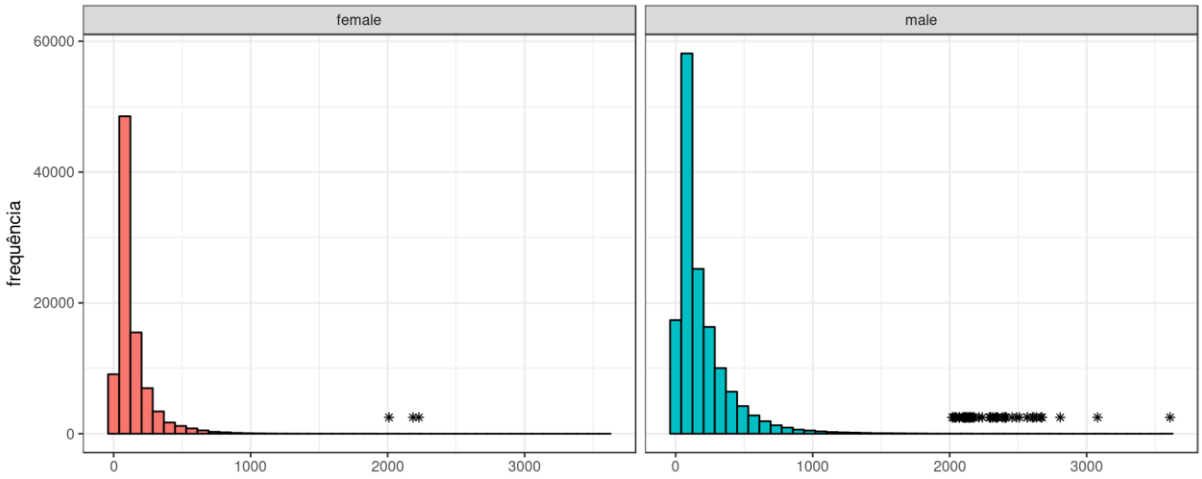


Figura 7: Histogramas da variável TI para medidas pré-tratamento de parasitas *S. mansoni*. Amostra: 88.527 medidas de fêmeas e 147.228 medidas de machos. Asteriscos indicam valores extremos, acima de 2000.

Uma abordagem comum na análise de dados não normais é o uso de transformações monótonas sobre as medidas (AITCHISON, 1982). Como as variáveis em estudo podem ser interpretadas em geral como proporções em função da sobreposição entre as duas imagens/objetos, a escolha natural de transformação a ser utilizada é a função *logit*, recomendada pela sua maior interpretabilidade[19] em

[19] O resultado da operação *logit* nada mais é que uma outra forma de expressar proporções. Seja *p* uma proporção: originalmente, *p* é limitado ao intervalo [0,1], enquanto que *logit*(*p*) repousa sobre a reta dos números reais. Se *p* for pensado em termos de probabilidade - assim como a proporção de, por exemplo, bolas pretas em uma urna pode ser pensada como a probabilidade de se sortear uma delas - *logit*(*p*) é uma medida da

termos de *log-odds* e grande popularidade, presente em muitos exemplos de estudos, inclusive de bioensaios (NORTON e DOWD, 2017). As figuras 8-10 são os histogramas para a variáveis FNR, CRI e CARI sob transformação *logit*.

Seria possível aplicar a mesma transformação para a variável TI reescalada (vide figura 3), mas se obteria essencialmente o mesmo resultado que se observa com a variável FNR na figura 8, dado o forte relacionamento entre estas duas variáveis que foi evidenciado pelos resultados de simulação. Como a variável TI é uma medida positiva (limitada somente devido a censura à direita), a transformação logarítmica foi empregada sobre esses dados. Os histogramas resultantes são apresentados na figura 11.

Todas as variáveis mostram histogramas com um perfil comum em forma de sino (ou neste caso, sinos), semelhantes à sobreposição de duas ou três gaussianas. Como a variável CRI não apresentou nenhuma característica em distribuição que a torne preferível às outras variáveis, nem a discussão anterior a respeito dos resultados de simulação lhe foi favorável, temos motivos mais que suficientes para eliminar esta variável da lista de candidatas. As variáveis FNR e CARI possuem perfis de distribuição mais semelhantes entre si, onde a principal diferença entre elas continua sendo somente a presença de valores nulos para FNR[20] e a possibilidade inconveniente de se obter valores de maiores que 1 para um índice (CARI) que fora construído para pertencer ao intervalo [0,1]. Dada a intuitiva importância dos valores nulos para a análise da motilidade e falta de qualquer indício nos dados que favoreça a escolha de CARI em detrimento de FNR, faz-se a opção por manter o estudo somente das variáveis FNR e TI.

Como visto pelos resultados de simulação, FNR é independente do tamanho do verme e TI é linearmente dependente à área que o verme efetivamente ocupa na imagem. Isto é suficiente para explicar a diferença na distribuição da variável FNR na escala *logit* (que se mostra mais concentrada em torno de um valor central da figura 8), e o perfil mais estratificado presente na distribuição da variável TI na escala logarítmica, em consequência do efeito da variação natural de tamanho que os vermes podem apresentar nesta amostra (que de fato é uma amostra

chance (*odds*, em inglês) de um evento em estudo ocorrer, expresso na escala logarítmica (NORTON e DOWD, 2017).

[20] Algumas medidas foram removidas somente para a confecção dos histogramas, como indicado nas legendas das figuras 8-11. Entretanto, estas medidas não foram de fato excluídas na modelagem dos dados.

estratificada não intencional decorrente da sua construção a partir de inúmeros experimentos independentes - vide figuras 12-13). A figura 12 mostra como os dados do experimento intitulado “x” implicaram a presença de um máximo local (abaixo de *logit* (FNR) < -5) presente na figura 8 para vermes fêmeas, enquanto que na figura 13 pode-se notar um período inicial onde os vermes utilizados são maiores e mais ativos e um segundo período onde eles são menores. Ademais, elas evidenciam que a medida de motilidade é sujeita a uma fonte de heterogeneidade ligada à identidade do verme, isto é, vermes diferentes podem apresentar níveis de atividade/motilidade sistematicamente diferentes uns dos outros.

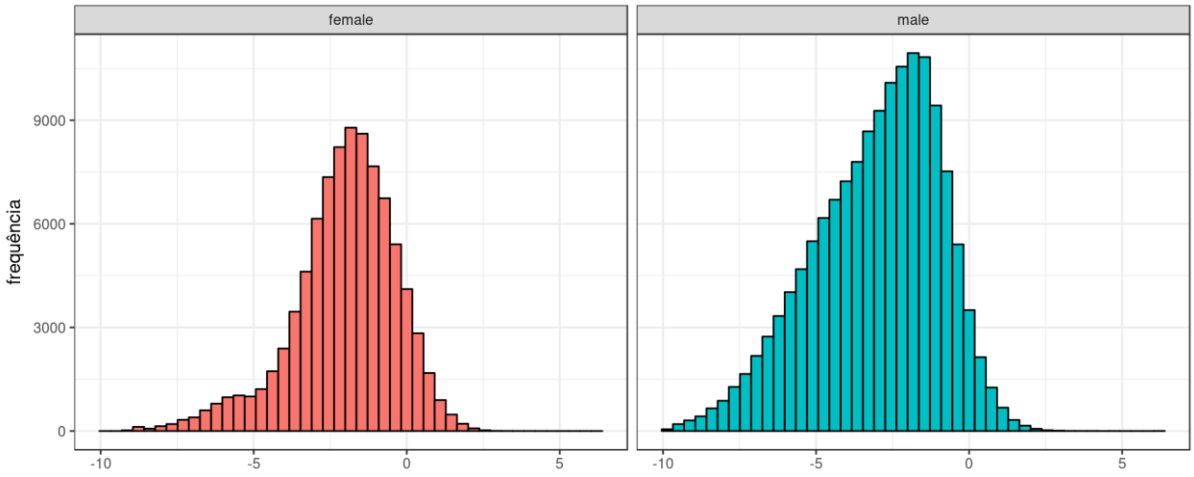


Figura 8: Histogramas da variável *logit* (FNR) para medidas pré-tratamento de parasitas *S. mansoni*. Amostra: 88.527 medidas de fêmeas e 147.228 medidas de machos. No total, 757 medidas foram removidas, sendo 16 devido FNR = 1 e 741 por FNR = 0.

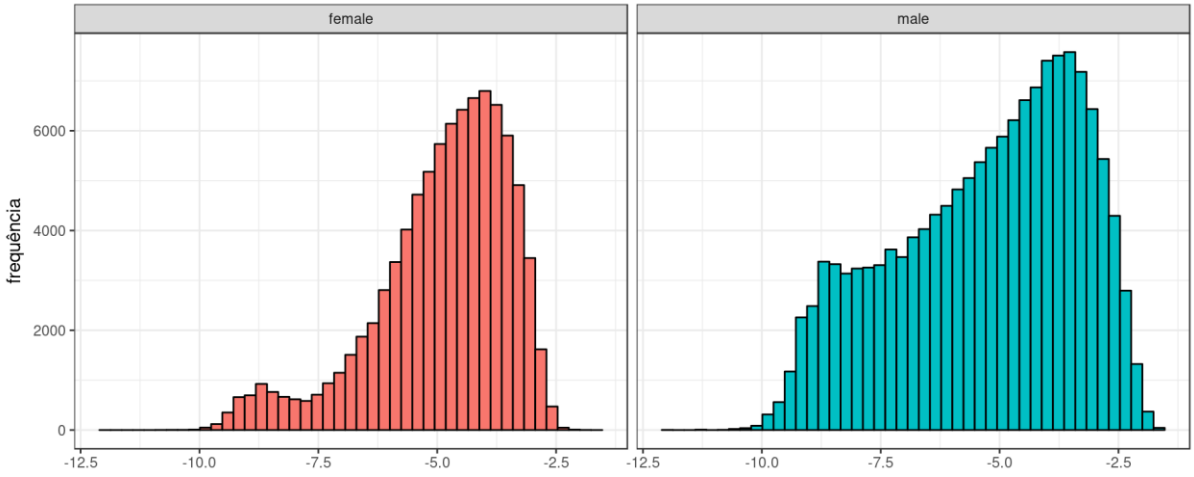


Figura 9: Histogramas da variável logit (CRI) para medidas pré-tratamento de parasitas *S. mansoni*. Amostra: 88.527 medidas de fêmeas e 147.228 medidas de machos.

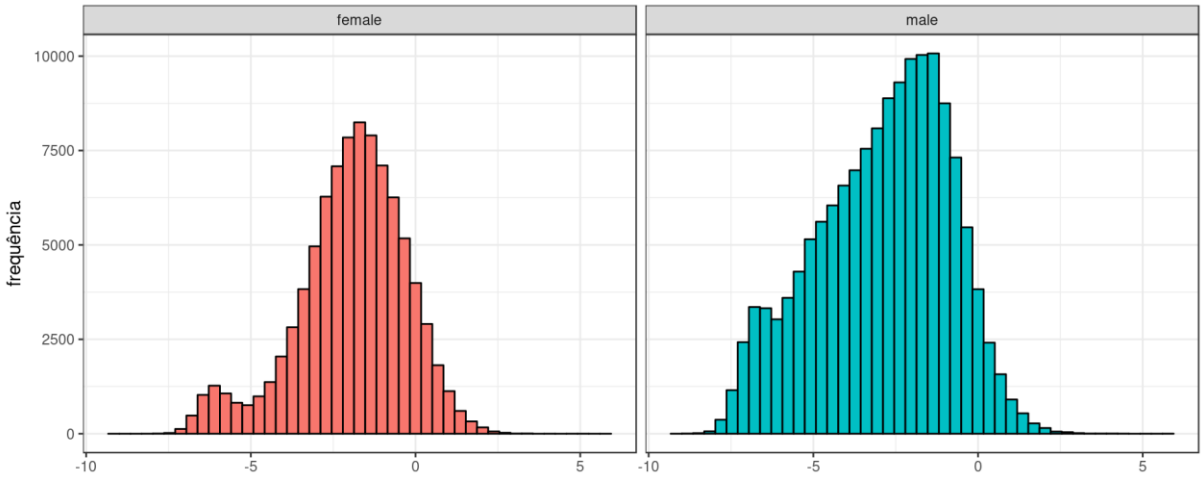


Figura 10: Histogramas da variável logit (CARI) para medidas pré-tratamento de parasitas *S. mansoni*. Amostra: 88.527 medidas de fêmeas e 147.228 medidas de machos. No total, 24 medidas foram removidas por CARI ≥ 1.

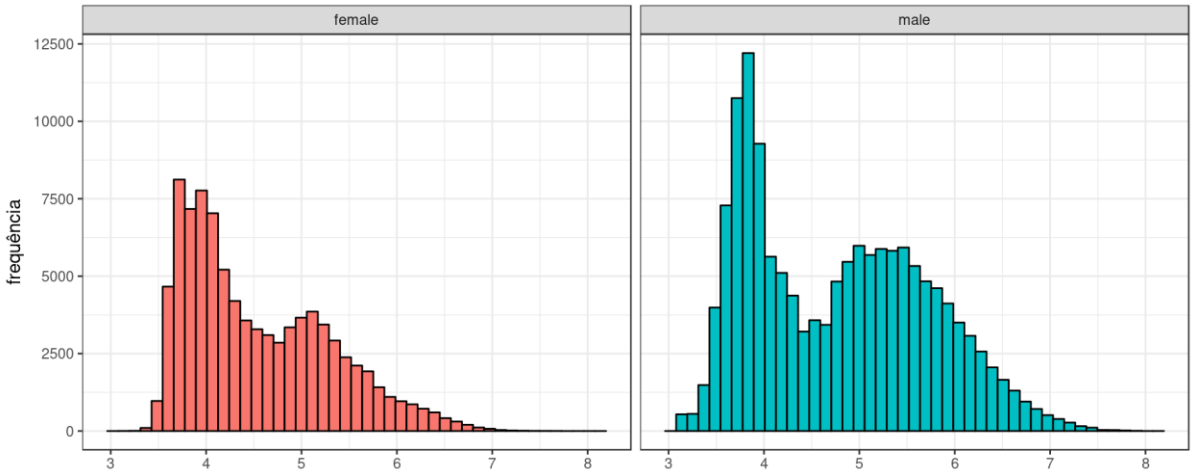


Figura 11: Histogramas da variável log (TI) para medidas pré-tratamento de parasitas *S. mansoni*. Amostra: 88.527 medidas de fêmeas e 147.228 medidas de machos.

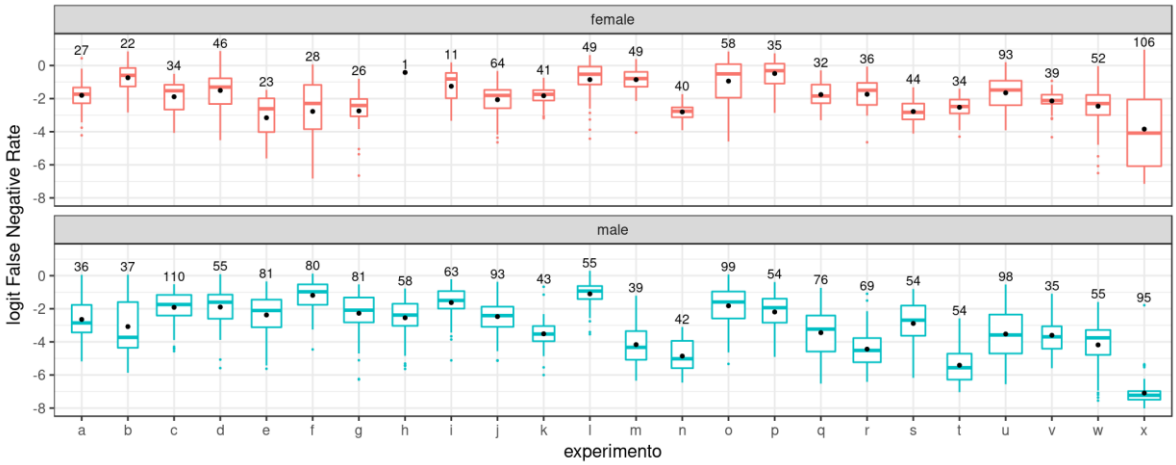


Figura 12: Boxplots da média de logit (FNR) por verme, para medidas pré-tratamento de parasitas *S. mansoni* para cada experimento independente desenvolvido no período 2014-2016, em ordem cronológica. No total, 88.527 medidas de fêmeas e 147.228 medidas de machos; 757 medidas foram removidas, sendo 16 devido FNR = 1 e 741 por FNR = 0. Os valores superescritos aos boxplots indicam o número de parasitas por experimento ao final do processamento dos dados.

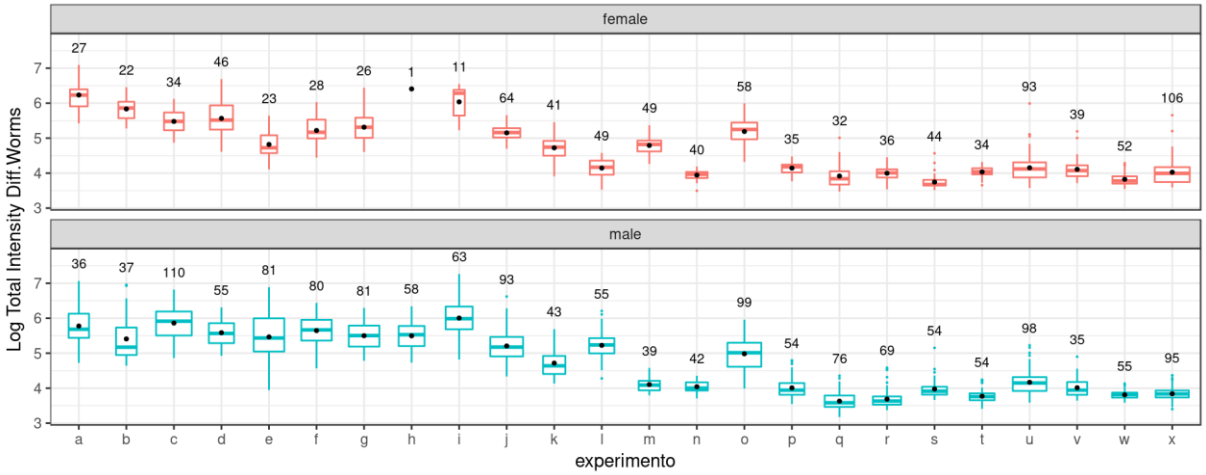


Figura 13: Boxplots da média de log (TI) por verme, para medidas pré-tratamento de parasitas *S. mansoni* para cada experimento independente desenvolvido no período 2014-2016, em ordem cronológica. No total, 88.527 medidas de fêmeas e 147.228 medidas de machos. Os valores superescritos aos boxplots indicam o número de parasitas por experimento ao final do processamento dos dados.

Outro ponto importante a ser avaliado é a extensão da autocorrelação, a princípio, presente nestes dados devido a estrutura sequencial dos dados obtidos por verme e por tempo que compõem o estudo. As figuras 14-15 mostram exemplos de dados seriais das variáveis FNR e TI transformadas, que correspondem a medidas pré-tratamento de uma amostra de 3 vermes, com o intuito de exemplificar

o comportamento observado. Os gráficos de dispersão apresentados sugerem que a autocorrelação entre as replicatas de medidas pré-tratamento de um verme é muito pequena, pois não se observa tendências de crescimento, de declínio ou periodicidade da medida de motilidade em função do tempo de medida. Como mostram as figuras 14-15, as 99 medidas de cada verme parecem estar dispersas aleatoriamente ao redor da média estimada. A distribuição das medidas, resumida pelos histogramas destas figuras, é, em geral, simétrica e não possui uma grande evidência contra a hipótese de normalidade. Este resultado sugere que os dados pré-tratamento de um mesmo verme podem ser tratados como aproximadamente independentes entre si, isto é, como replicatas da medida de motilidade.

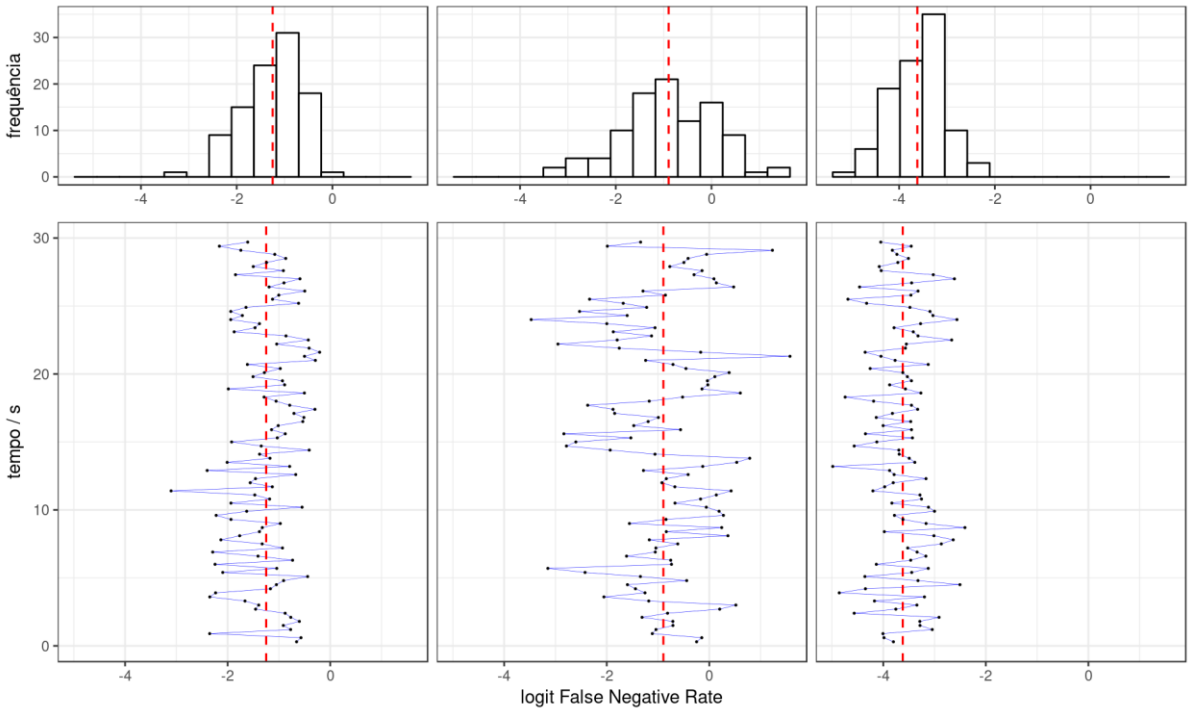


Figura 14: Distribuição das medidas (replicatas) logit (FNR) para uma amostra exemplo de 3 vermes. A parte superior resume a informação das replicatas de cada verme em histogramas; a parte inferior mostra em detalhe as 99 medidas individuais de um ensaio, por meio de gráficos de dispersão. A linhas vermelhas pontilhadas indicam o valor médio das 99 replicatas de cada verme.

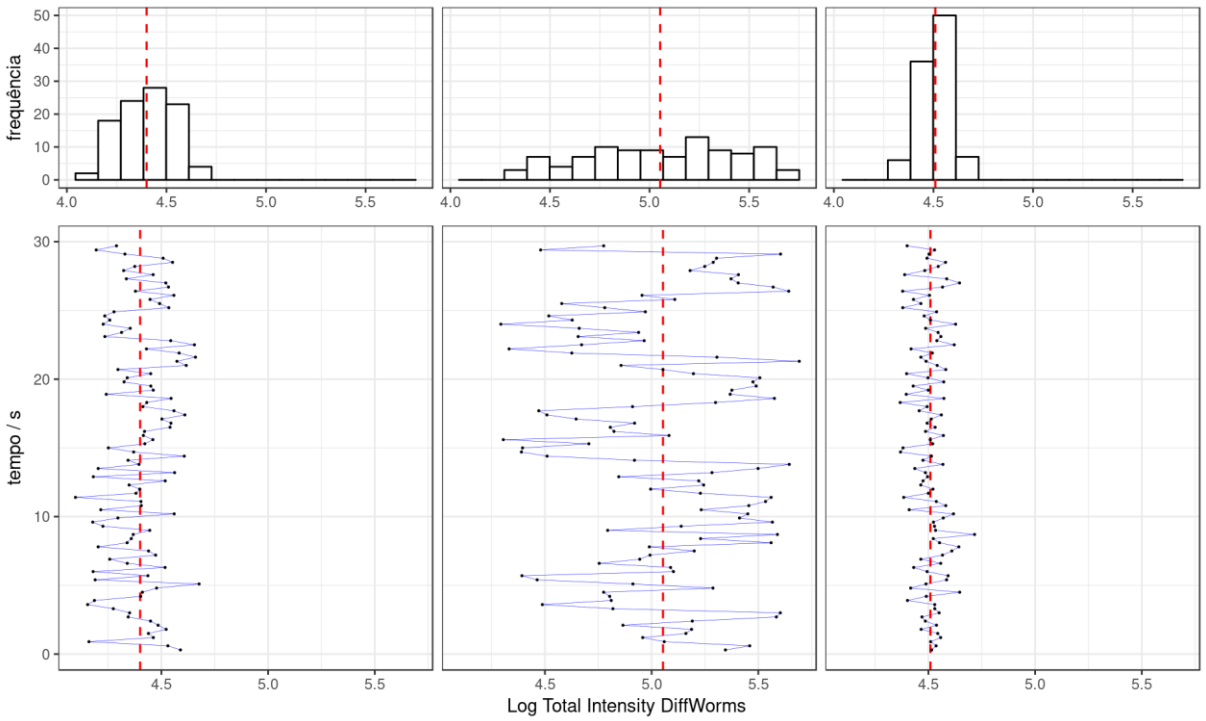


Figura 15: Distribuição das medidas (replicatas) log (TI) para uma amostra exemplo de 3 vermes. A parte superior resume a informação das replicatas de cada verme em histogramas; a parte inferior mostra em detalhe as 99 medidas individuais de um ensaio, por meio de gráficos de dispersão. A linhas vermelhas pontilhadas indicam o valor médio das 99 replicatas de cada verme.

A figura 16 ilustra as medidas FNR na escala *logit* de outros três vermes do grupo controle ao longo de todo o experimento, mostrando as replicatas e a média estimada por verme e por tempo do ensaio (pré-tratamento, 0h, 24h, 48h e 72h). Ela ajuda a reforçar a hipótese de autocorrelação baixa entre as replicatas e vai além, mostrando a existência de uma pequena flutuação das médias estimadas de *logit* (FNR) para cada tempo de ensaio. Esta é uma característica importante de ser notada, porque dela pode-se inferir uma possível fonte de heterogeneidade na medida de motilidade, associada à mudança natural no comportamento de um parasita ao longo do tempo que afete seu movimento. Como a duração de um ensaio é cerca de 30 segundos por verme, é improvável que ocorram mudanças notáveis dentro deste período, mas pode-se notar mais facilmente mudanças entre um ensaio e outro, que são separados por muitas horas, como mostram as figuras 16-17. Os dados da segunda linha da figura 16 esboçam um exemplo nítido desta consideração. Deve-se ter em mente, apesar, que isto pode ser resultado da pequena autocorrelação devido a estrutura sequencial dos dados.

A figura 17 apresenta resultados semelhantes para a variável *log*(TI), somados a fortes sinais de que o grau de dispersão entre replicatas não é homogêneo ao redor da média estimada. Para a variável *logit*(FNR) pode existir algum grau de heterocedasticidade, mas em uma extensão muito menor que para a variável *log*(TI), como pode-se verificar pela análise qualitativa das figuras 16 e 17.

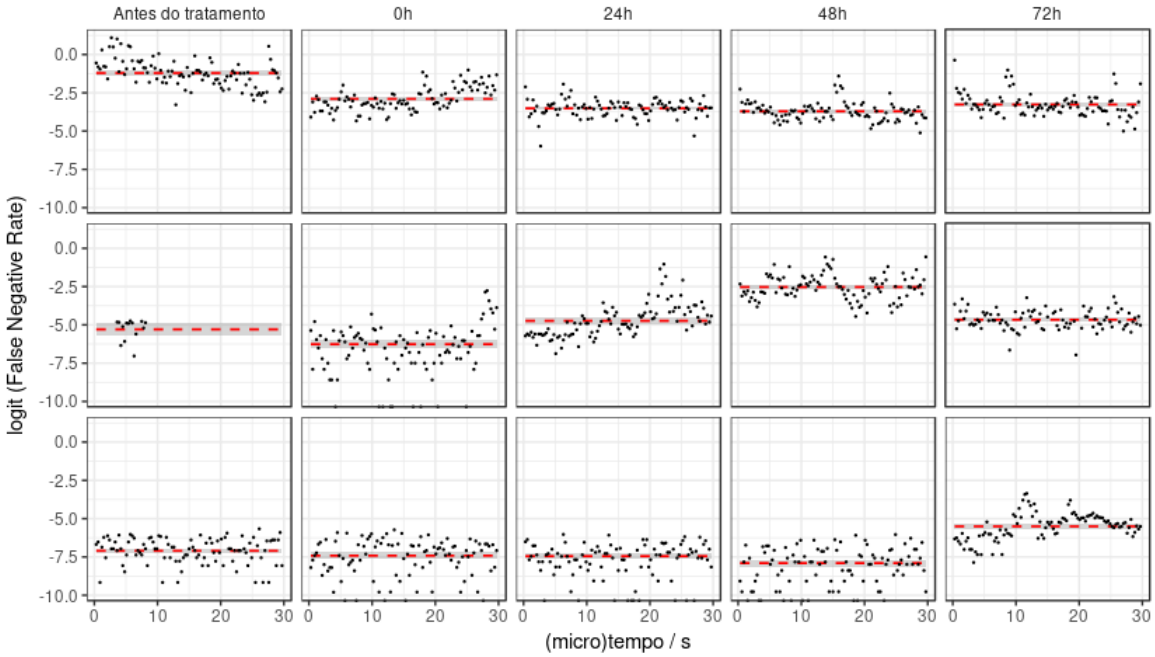


Figura 16: Conjunto de gráficos de dispersão das replicatas da medida logit (FNR) vs. o (micro) tempo para 3 vermes do grupo controle. Nesta tabela, cada linha contém dados de um único parasita e as colunas indicam o tempo do ensaio, em horas. O marco zero no (micro) tempo é relativo ao início de cada ensaio. A linha vermelha indica a média estimada; e a região cinza indica o IC 95% da estimativa.

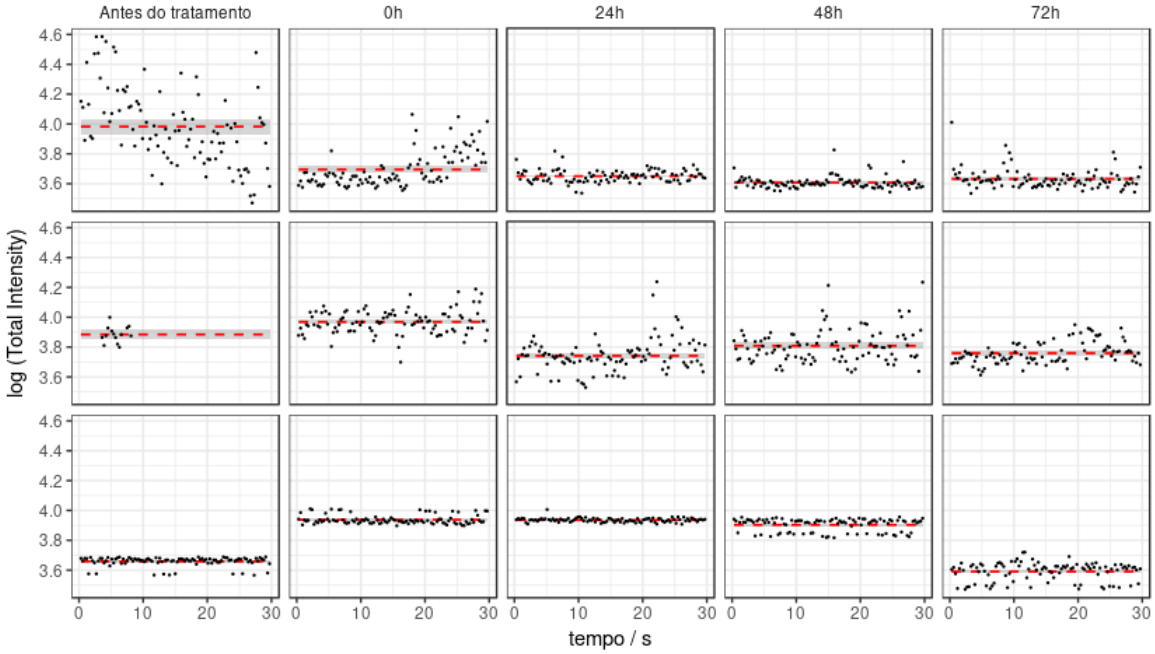


Figura 17: Conjunto de gráficos de dispersão das replicatas da medida log (TI) vs. o (micro) tempo para três vermes do grupo controle. Nesta tabela, cada linha contém dados de um único parasita e as colunas indicam o tempo do ensaio, em horas. O marco zero no (micro) tempo é relativo ao início de cada ensaio. A linha vermelha indica a média estimada; e a região cinza indica o IC 95% da estimativa.

Ainda pela análise das figuras 16 e 17, pode-se notar que as transformações realizadas nas variáveis não conseguiram estabilizar a variância das medidas por verme, apesar de ter alcançado um formato simétrico para a distribuição (figuras 14 e 15). De fato, espera-se que vermes enfraquecidos (por terem sido tratados com um composto ativo, por exemplo) exibam, em média, baixa motilidade, mas também um baixo desvio padrão em torno desta motilidade média; o caso limite é atingido quando a motilidade é zero e a variância também é nula. Estes resultados e discussão sugerem que um modelo linear, mesmo usando as variáveis transformadas, não é recomendável em consequência da variância não uniforme em relação à média estimada. Além do que, as transformações inutilizam medidas de valor zero, possíveis em ambas as variáveis, e que são muito relevantes na análise para serem simplesmente desprezadas. Ao invés disso, é interessante o uso de MGLMs que têm a capacidade de acomodar dados onde a variância é uma função da média (GELMAN e HILL, 2007).

A fim de propor um MGLM para o ajuste de dados de motilidade, é de grande utilidade estudar a relação entre média e variância, ou equivalentemente a relação entre a média e o desvio-padrão. Para isso, estimou-se a média e a variância das

medidas pré-tratamento de motilidade, agrupadas por verme. Entretanto, como a variável TI é extensiva, pois é dependente da área do verme, e FNR é intensiva, a comparação entre as duas necessita da conversão de uma das duas variáveis. Por isso, foi gerada uma nova variável extensiva pelo produto entre a área do verme (isto é, o número total de *pixels* - dado pela variável *Área*) e FNR (a proporção de *pixels* classificados como falso negativos), que foi chamada de FN (representando *o número* de *pixels* classificados como falso negativos).

A análise qualitativa da relação funcional entre média e variância das variáveis FN e TI, ambas extensivas e linearmente dependentes à área do verme, é ilustrada nas figuras 18-19. Observa-se por meio destes gráficos log-log relações aproximadamente lineares entre médias e desvios-padrão. Na figura 18 é evidente uma relação linear para valores moderados de motilidade FN média, porém vermes com motilidades FN elevadas geram estimativas de variância inferiores àquelas esperadas pela tendência linear inicialmente observada. Acredita-se que isto é consequência da censura à direita, que é uma característica destas variáveis, como discutido anteriormente. Nota-se também que a relação $\log(\sigma) \times \log(\mu)$ é essencialmente 1:1, independentemente do sexo do parasita, o que equivale afirmar que a relação entre média e variância é de segunda ordem.

Enquanto isso, a variável TI exibe uma relação média × variância também linear em escalas logarítmicas (figura 19), mas, ao contrário da variável FN, os desvios à linearidade são muito acentuados e ocorre com os vermes para os quais os valores de motilidade TI média são baixos. Notável que - apesar de serem conceitualmente relacionadas e consideradas quase equivalentes na análise da motilidade por meio de dados simulados - a complexidade experimental e de processamento das imagens foram suficientes para afetar a resposta destas variáveis de forma tão marcante.

Destes últimos resultados, constata-se que a variável FN é a melhor candidata para representar a motilidade dentre os tipos de medida disponíveis. Além disso, os resultados desta análise exploratória servem de base para o desenvolvimento de MGLMs em busca de descrever os dados de motilidade, introduzidos na próxima seção.

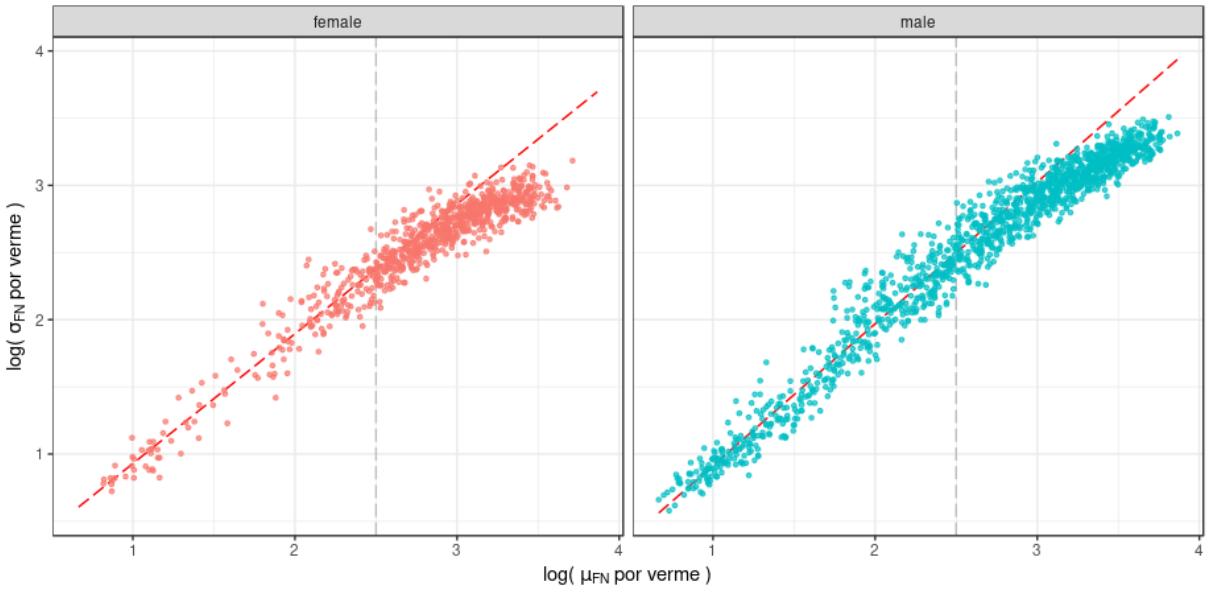


Figura 18: Gráficos de dispersão da relação entre a média e o desvio padrão das medidas FN agrupadas por verme - escala logarítmica (base 10) - usando-se somente medidas pré-tratamento. No total, há dados de 822 vermes fêmeas e 1432 vermes machos. As linhas tracejadas vermelhas indicam as retas ajustadas com o uso de dados de vermes com motilidade média inferior a 102,5.

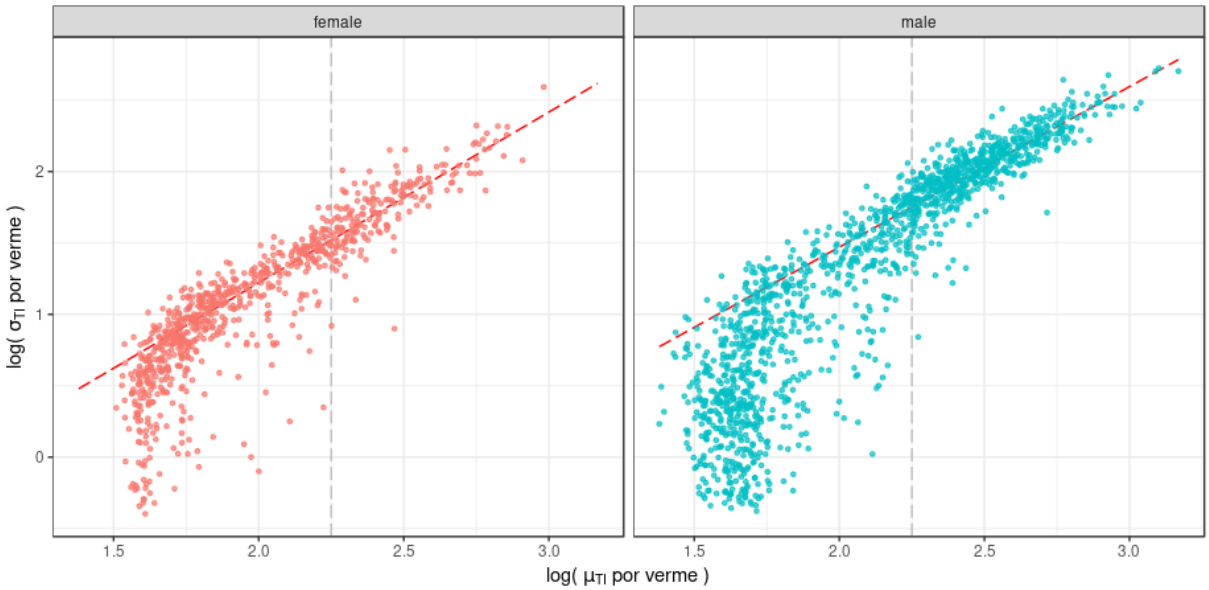


Figura 19: Gráficos de dispersão da relação entre a média e o desvio padrão das medidas TI agrupadas por verme - escala logarítmica (base 10) - usando-se somente medidas pré-tratamento. No total, há dados de 822 vermes fêmeas e 1432 vermes machos. As linhas tracejadas vermelhas indicam as retas ajustadas com o uso de dados de vermes com motilidade média superior a 102,25.

4.2 Proposição e comparação de modelos estatísticos para a motilidade

Para levar em conta a incerteza na medida de motilidade, escolheu-se tratá-la como uma variável aleatória contável[21], chamada FN, estudar a aplicação de algumas distribuições de probabilidade adequadas para este tipo de dado e diagnosticar o ajuste por meio da análise dos resíduos. Enquanto que no estudo de variáveis contínuas a distribuição normal é a referência, para variáveis contáveis a distribuição de Poisson é geralmente o ponto de partida.

A medida da motilidade é dada pelo número de eventos que ocorrem em um intervalo de tempo fixo (300 ms), onde o "evento" é a classificação de um *pixel* da Im_2 como falso-negativo, indicando a ocorrência de um movimento do parasita. Entretanto, é esperado que tais eventos possam ser simultâneos, pois ocorre o movimento conjunto de muitos segmentos do parasito praticamente ao mesmo tempo. Daí já se pode esperar que uma regressão Poisson não seria adequada para os dados de motilidade, pois ela assume que eventos simultâneos não ocorrem. A figura 18 mostra que a evidência experimental também corrobora esta assertiva, pois experimentalmente os dados pré-tratamento indicam que a relação entre média e variância é de segunda ordem (diferentemente da distribuição Poisson).

Neste trabalho foram testadas cinco distribuições de probabilidade que poderiam ser úteis na modelagem dos dados de motilidade dos parasitas:

a) Beta-binomial;
b) Beta-binomial inflada por zeros;
c) Binomial Negativa;
d) Binomial Negativa inflada por zeros;
e) Poisson Generalizada.

[21] Antes de explicar os modelos testados, vamos esclarecer a definição operacional da motilidade dada pela variável FN. Como todas as variáveis anteriormente estudadas, uma medida False Negative (FN) é resultado da comparação entre os *pixels* de duas bioimagens sequenciais. Sejam duas imagens consecutivas chamadas Im_1 e Im_2. Para cada uma delas, existe um conjunto de *pixels* classificados como *pixels* de primeiro plano, que identificam o "objeto" verme. A variável FN é definida como o número de *pixels* de primeiro plano em Im_1 que não pertencem ao primeiro plano de Im_2. Em outras palavras, ela é o número de *pixels* que constatam a ocorrência de movimento no intervalo de tempo entre a captura das duas imagens, e dá uma medida indireta do deslocamento (nos termos da figura 2). Logo, a motilidade segue como uma medida do tipo contagem, isto é, somente assume valores inteiros não-negativos.

Como citado anteriormente, estas distribuições são todas utilizadas para modelagem de dados do tipo contagem e foram aplicadas aqui em busca de capturar características observadas na análise exploratória. A distribuição beta-binomial e sua versão inflada por zeros foram testadas para verificar se poderiam modelar adequadamente a censura à direita característica da medida de motilidade FN. A função-link usada para a ligação entre média μ e preditor linear η foi a função *logit*, aplicada a μ_p. Isto é equivalente a relação

$$\mu_p = \frac{\exp(\eta)}{1 + \exp(\eta)}$$

A análise exploratória mostrou que a função *logit* foi capaz de normalizar as medidas FNR. Esta variável é uma medida da *proporção* de *pixels* "indicadores de movimento", então pode ser interpretada como uma estimativa da probabilidade de um "evento" (a classificação de um *pixel* como falso-negativo) ocorrer. A mesma interpretação é dada a μ_p , o que justifica a escolha da função-*link logit*. O parâmetro *n*, que limita superiormente a distribuição, foi feito igual a variável *Área*, já que ela indica o valor máximo para a medida de motilidade FN (vide tabela 1).

As distribuições Poisson Generalizada, binomial negativa e sua versão inflada por zeros não podem levar a presença de censura à direita em consideração. Considerando-se, entretanto, que esta limitação na mensuração da motilidade possa ser minimizada e que ela é uma característica menos importante da medida de motilidade FN, as outras distribuições podem ser úteis. Uma justificativa teórica para testar o modelo de regressão BN é, que, na sua parametrização, está implicitamente incorporada[22] a função *logit*, e considerações análogas ao caso beta-binomial são válidas também.

A distribuição Poisson Generalizada[23] é uma opção semelhante à distribuição binomial negativa, pois ela também é reconhecida como uma distribuição de mistura de Poisson (JOE e ZHU, 2005). Usou-se a parametrização GP-P com o parâmetro p = 1,5 para estar de acordo com o resultado experimental da figura 18 (isto é, uma relação de segunda ordem entre média e variância).

O modelo linear generalizado se torna multinível quando as fontes de heterogeneidade são inclusas no preditor linear como efeitos aleatórios, isto é,

[22] Considerando-se o uso da função-*link* logarítmica, como é feito neste trabalho.

[23] A distribuição GP inflada por zeros não foi testada, pois esta distribuição ainda não foi implementada no *package* R-INLA.

coeficientes, que variam ao nível de grupo, definidos como realizações individuais de variáveis aleatórias (GELMAN e HILL, 2007). Da análise exploratória, resultaram duas possíveis fontes de heterogeneidade independentes do tratamento a serem avaliadas: a primeira a nível de verme, e a outra dada pela interação verme×tempo. Tomadas conjuntamente, definem o efeito aleatório δ a nível de verme. Além deste, é preciso considerar o efeito do tratamento *β*, que é função do composto e do tempo. Com isso, o preditor linear η_i (para a *i*-ésima medida) dos modelos propostos é constituído pela soma de um efeito fixo e 2 efeitos aleatórios:

$$\eta_i = \alpha + \beta_{(c,t)[i]} x_i + \delta_{(w,t)[i]} \quad \begin{cases} x_i = 0\,, & se\ t[i] = 1 \\ x_i = 1\,, & c.c \end{cases}$$

onde os índices de grupo são: *c* indicando tratamentos (compostos), com *c*=1 representando o grupo controle; *w* para os vermes (*worms*); e *t* para o tempo do ensaio/leitura, sendo *t*=1 indicador de dados pré-tratamento, enquanto os demais tempos são indicados pelos números inteiros maiores do que ou igual a dois.

4.2.1. Modelos para experimentos tipo triagem de compostos químicos

Foram testados três tipos de efeito a nível de verme e três modelos para o efeito de tratamento, como descrito anteriormente. Para o efeito de tratamento, a análise da estrutura química gerou uma rede esparsa, formando cinco ilhas para o modelo usando a rede de *scaffolds.* Para o conjunto de dados estudado, se obteve um único *scaffold* compartilhado por três compostos: Artemeter, Artesunato e Dihidro-artemisinina (Figura 20).

O modelo baseado em similaridade Tanimoto resultou em uma rede totalmente conectada (Figura 21), já que a matriz de adjacência usada é ponderada pelos coeficientes de similaridade Tanimoto. A menor similaridade não nula obtida foi de 1,9% (entre o Clonazepan e o Artesunato) e a maior similaridade foi de 74,7% entre Artemeter e Dihidro-artemisinina. Nota-se, entretanto, que esta abordagem teve como resultado principal a detecção da grande similaridade entre o trio de compostos Artemeter, Artesunato e Dihidro-artemisinina. Ou seja, as duas abordagens obtiveram o mesmo resultado qualitativo: o trio de compostos supracitado formam um grupo, enquanto os demais compostos tem essencialmente nenhuma similaridade um com os outros.

Avaliou-se o modelo em três versões para cada distribuição de probabilidade e para cada sexo separadamente, como detalhado anteriormente. Quando são comparados os efeitos estimados da distribuição beta-binomial com os da sua versão inflada por zeros, verifica-se que o parâmetro α (que governa o grau de inflação por zeros) foi estimado essencialmente nulo[24], isto é, estimou-se que o grau de inflação é desprezível. Consequentemente, os demais coeficientes e hiperparâmetros estimados foram essencialmente idênticos entre si, quando se compara os resultados das duas distribuições[25].

Uma análise dos resíduos escalados [26] foi realizada para estudar a adequação dos diferentes modelos aos dados. As figuras 22-25 apresentam gráficos quantil-quantil, de resíduos escalados, cuja distribuição teórica é a distribuição uniforme; enquanto as figuras 26-29 mostram os gráficos de resíduos escalados em função do preditor linear[27].

[24] Para as fêmeas, a distribuição *a posteriori* do parâmetro α apresentou mediana = $7{,}3 \cdot 10^{-6}$ e desvio padrão = $1{,}9 \cdot 10^{-5}$. Para os vermes machos, a mediana = $5{,}2 \cdot 10^{-6}$ e o desvio padrão = $1{,}3 \cdot 10^{-5}$. (Estes resultados são especificamente oriundos do modelo SEM usando a rede de *scaffolds*, mas os outros modelos renderam resultados muito similares).

[25] Por este motivo, as demais comparações apresentadas neste trabalho não incluem a distribuição beta-binomial inflada por zeros.

[26] Obtidos pelo método de extração de resíduos escalados através de simulações da distribuição preditiva *a posteriori* (HARTIG, 2018). Para fins comparativos, uma análise visual por meio dos gráficos citados é suficiente para a discussão inicial dos resultados.

[27] Por simplicidade, somente são ilustrados os resultados obtidos pela abordagem com uso da rede de *scaffolds*, já que os resultados usando a abordagem por similaridade Tanimoto foram essencialmente iguais.

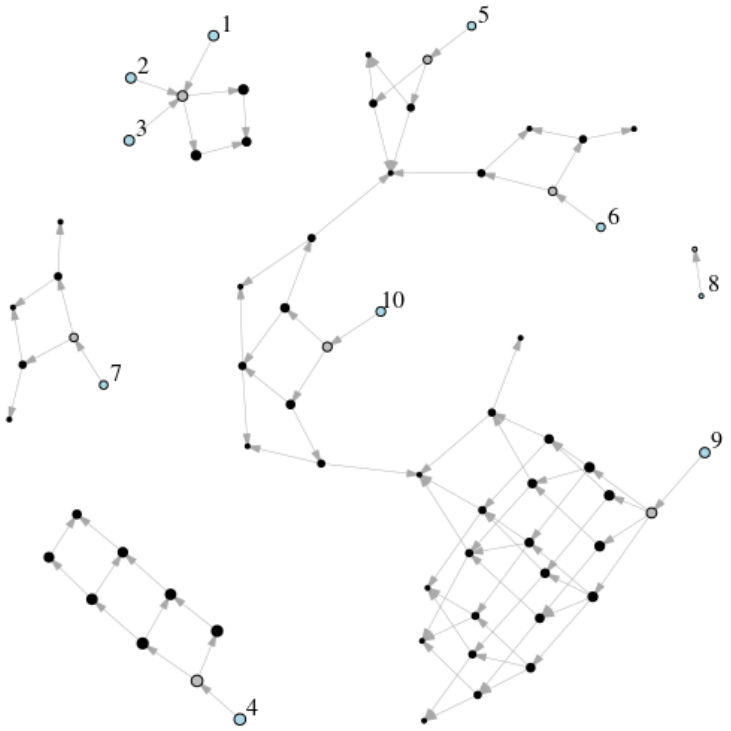


Figura 20: Rede completa de compostos, *scaffolds* e *sub-scaffolds*. Círculos azuis: compostos; círculos cinzas: *scaffolds*; círculos pretos: *sub-scaffolds*. 1) Artemeter; 2) Artesunato; 3) Dihidro-artemisinina; 4) Ácido gambógico; 5) Clonazepan; 6) Mefloquina; 7) Anfotericina B; 8) Orlistat; 9) Picrotoxina; 10) Praziquantel.

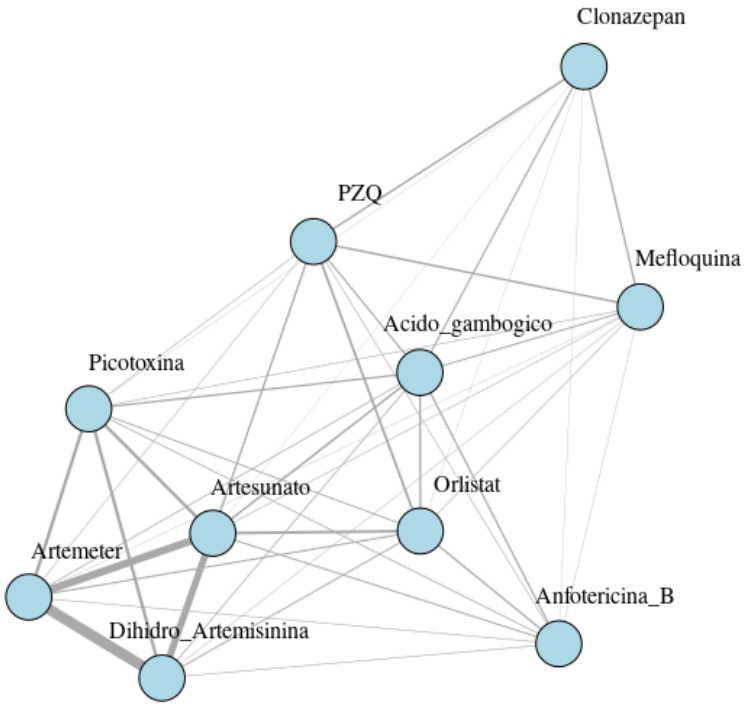


Figura 21: Rede de similaridade Tanimoto. A largura das arestas é proporcional ao grau de similaridade.

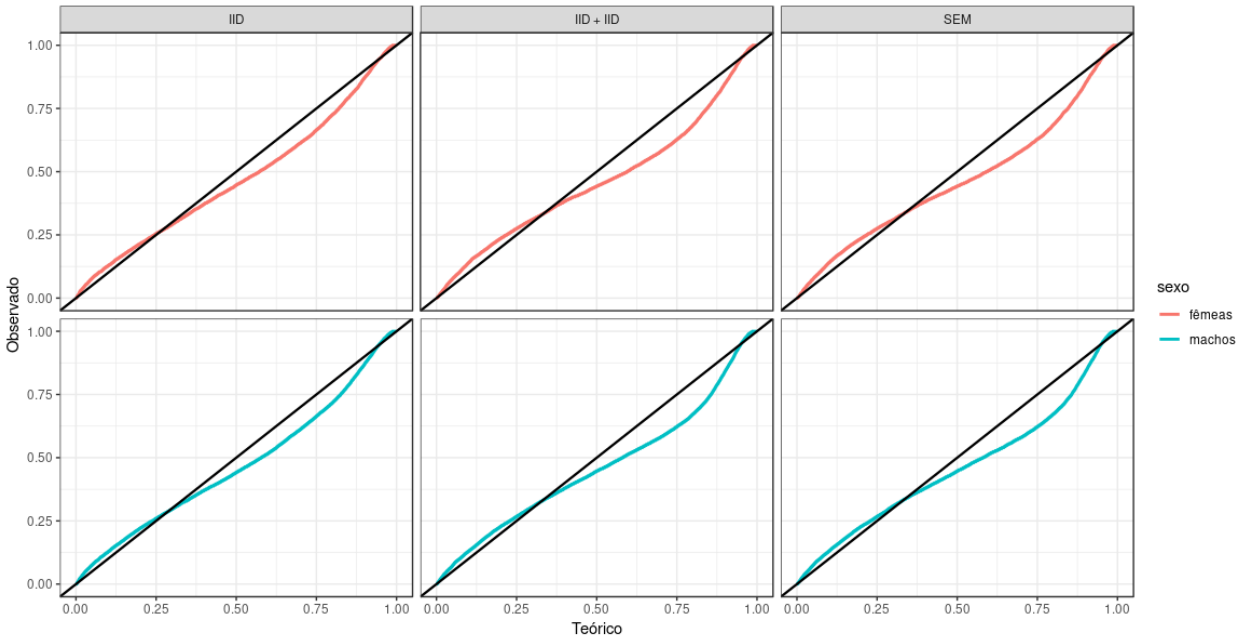


Figura 22: Gráficos quantil-quantil dos resíduos escalados contra a distribuição uniforme padrão, para a distribuição Beta-Binomial.

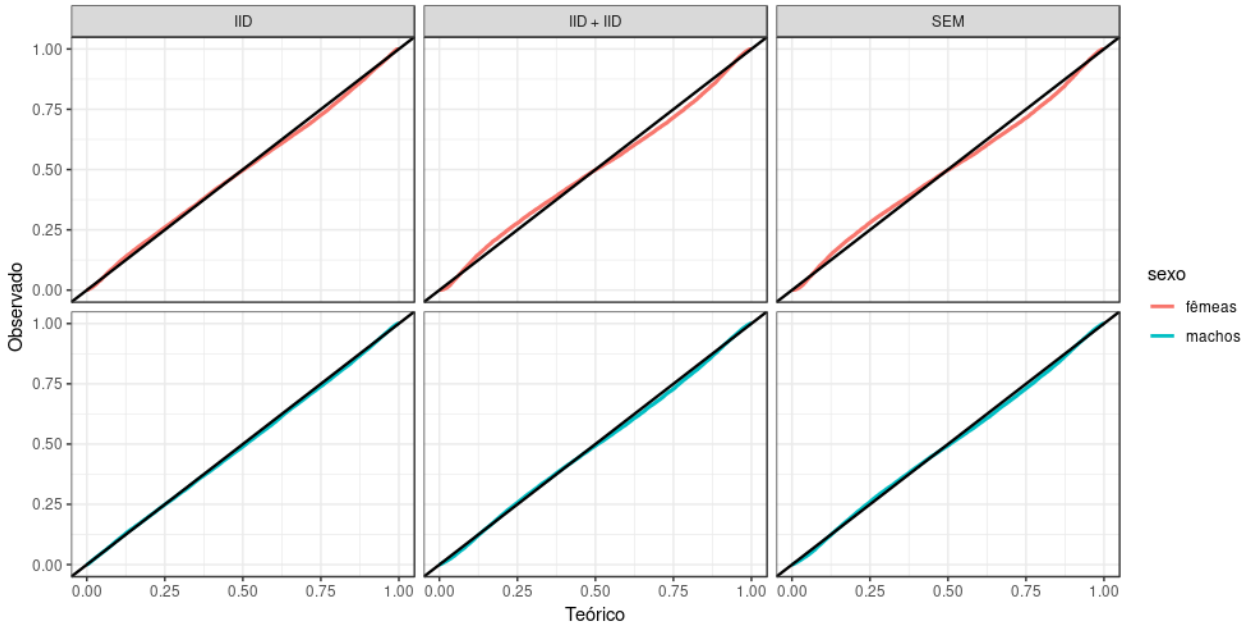


Figura 23: Gráficos quantil-quantil dos resíduos escalados contra a distribuição uniforme padrão, para a distribuição Binomial Negativa.

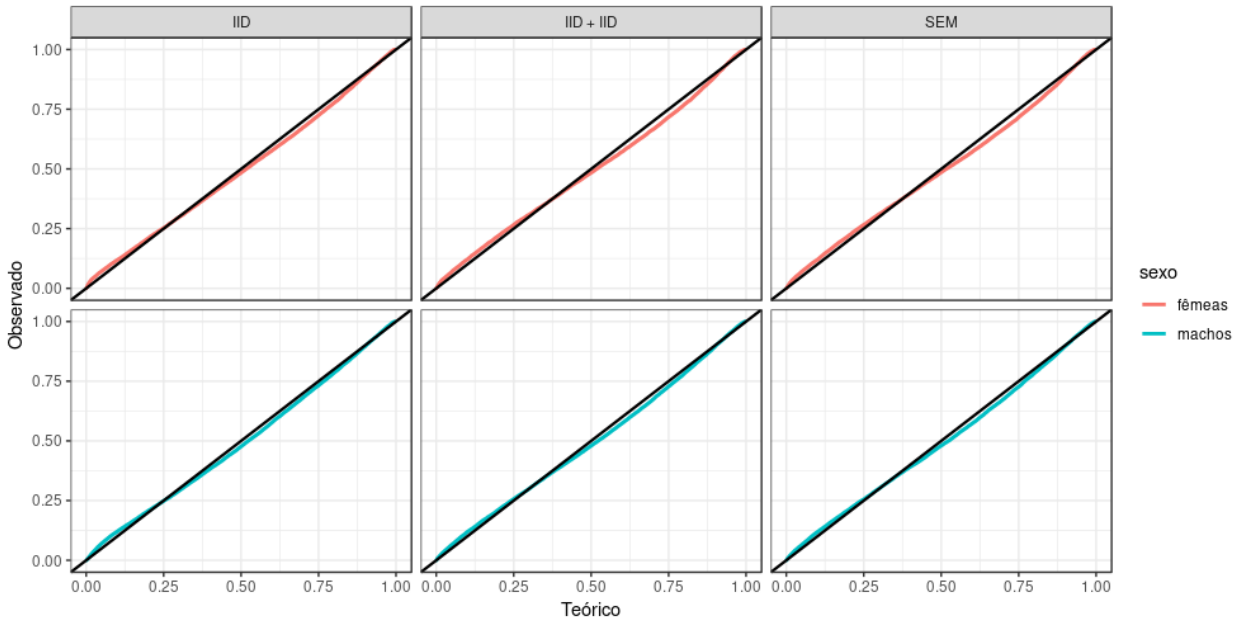


Figura 24: Gráficos quantil-quantil dos resíduos escalados contra a distribuição uniforme padrão, para a distribuição Binomial Negativa inflada por zeros.

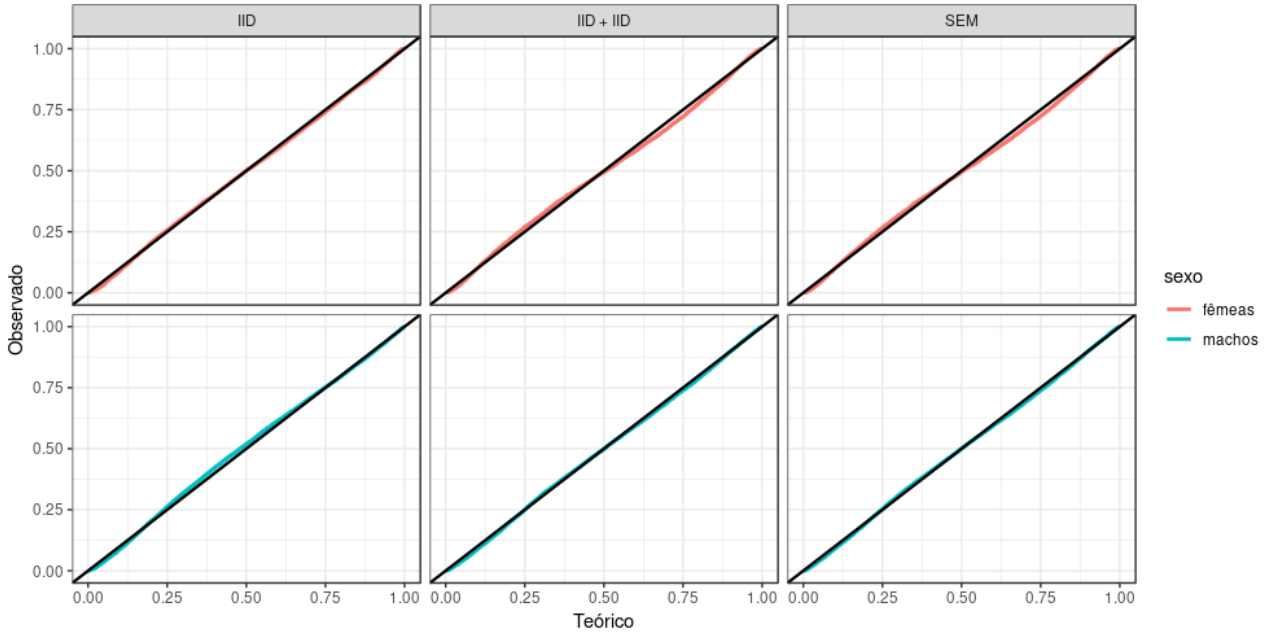


Figura 25: Gráficos quantil-quantil dos resíduos escalados contra a distribuição uniforme padrão, para a distribuição Poisson Generalizada.

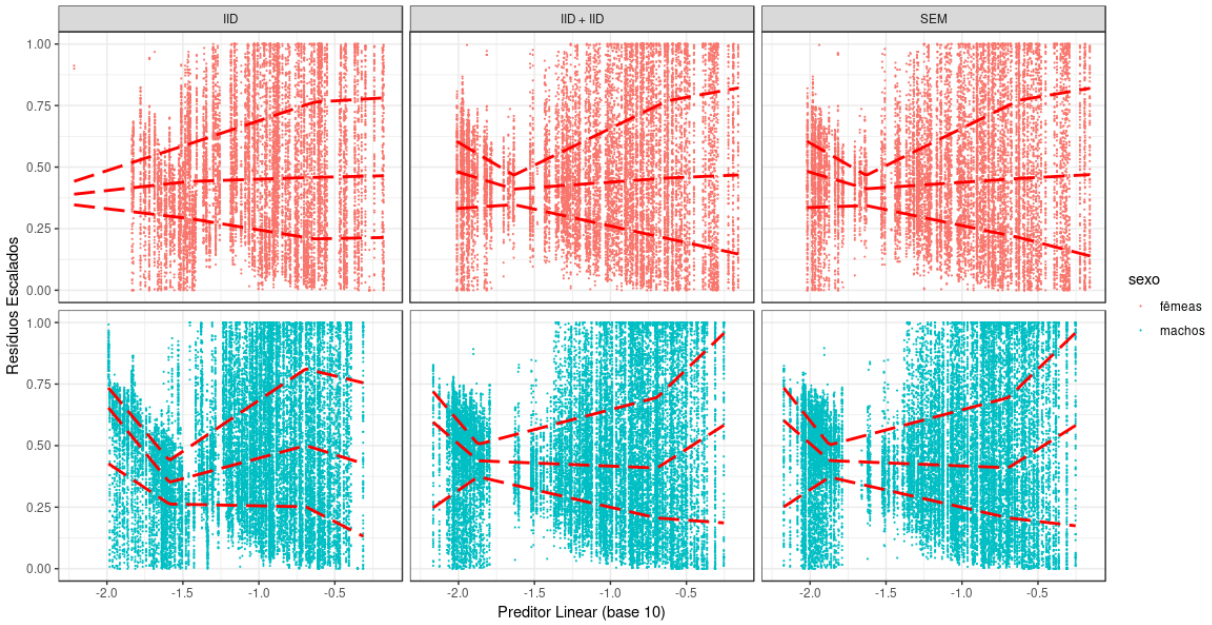


Figura 26: Gráficos de resíduos escalados (da distribuição Beta-Binomial) vs. preditor linear na base 10, mostrando a diferença entre os modelos para o efeito a nível de verme. Os resíduos são esperados apresentar distribuição uniforme padrão ao longo do eixo vertical. As linhas vermelhas tracejadas são retas estimadas por regressão quantil dos quantis 0,25, 0,5 e 0,75.

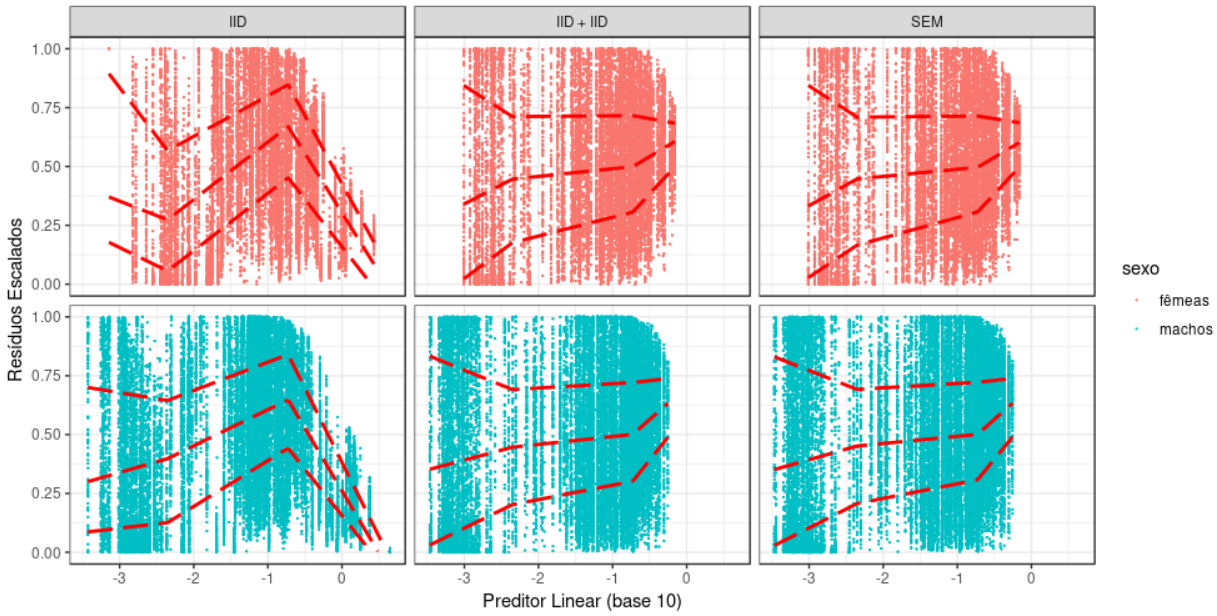


Figura 27: Gráficos de resíduos escalados (da distribuição Binomial Negativa) vs. preditor linear na base 10, mostrando a diferença entre os modelos para o efeito a nível de verme. Os resíduos são esperados apresentar distribuição uniforme padrão ao longo do eixo vertical. As linhas vermelhas tracejadas são retas estimadas por regressão quantil dos quantis 0,25, 0,5 e 0,75.

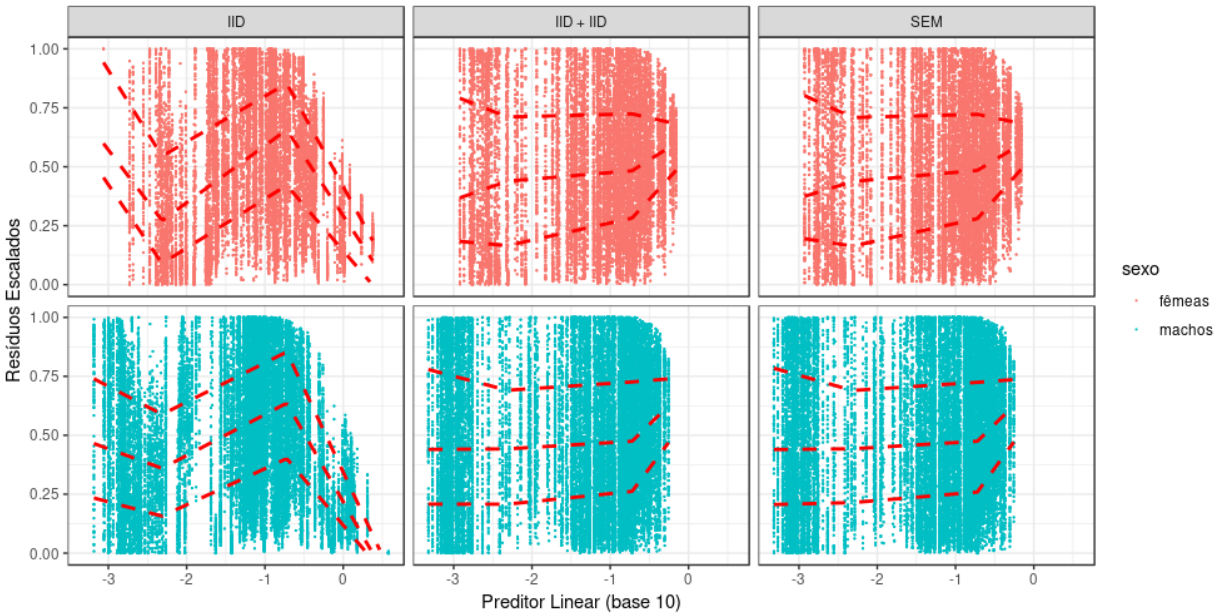


Figura 28: Gráficos de resíduos escalados (da distribuição Binomial Negativa inflada por zeros) vs. preditor linear na base 10, mostrando a diferença entre os modelos para o efeito a nível de verme. Os resíduos são esperados apresentar distribuição uniforme padrão ao longo do eixo vertical. As linhas vermelhas tracejadas são retas estimadas por regressão quantil dos quantis 0,25, 0,5 e 0,75.

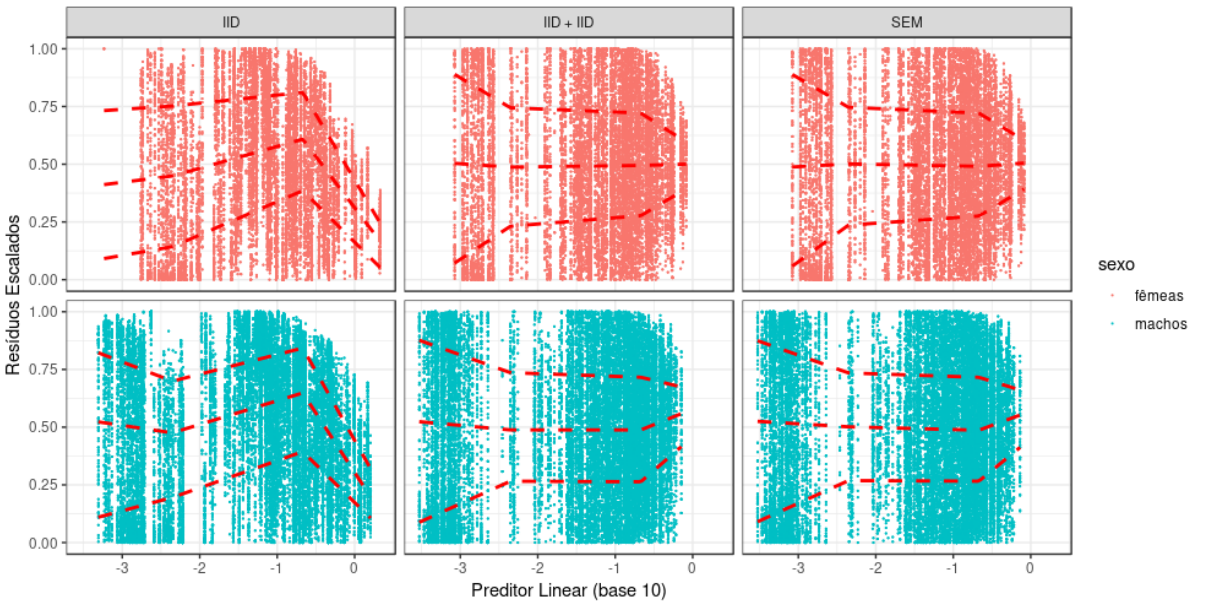


Figura 29: Gráficos de resíduos escalados da distribuição Poisson Generalizada vs. preditor linear na base 10, mostrando a diferença entre os modelos para o efeito a nível de verme. Os resíduos são esperados apresentar distribuição uniforme padrão ao longo do eixo vertical. As linhas vermelhas tracejadas são retas estimadas por regressão quantil dos quantis 0,25, 0,5 e 0,75.

A figura 22 permite notar que a distribuição beta-binomial não apresenta um bom ajuste geral aos dados, enquanto as figuras 23-25 apresentam graficamente

que as distribuições que generalizam a distribuição Poisson têm um desempenho muito superior, especialmente no ajuste dos dados de vermes machos. Estes resultados de ajuste geral são independentes do modelo a nível de verme empregado, indicando que a distribuição beta-binomial em si, ou mesmo sua extensão inflada por zeros, não são adequadas para modelar os dados de motilidade.

Os gráficos de resíduos escalados (figuras 26-29) permitem avaliar o ajuste em função do grau de motilidade[28], em termos do preditor linear η. As linhas vermelhas tracejadas ilustram os ajustes lineares para avaliar as tendências observadas em três níveis de motilidade separadamente (baixa, mediana e alta motilidade)[29] para os 1º, 2º e 3º quartil. A figura 26 corrobora a inadequação da distribuição beta-binomial aos dados. Em especial, nota-se que o ajuste para motilidades baixas e medianas superestima a variância dos dados e é dominada pelo ajuste dos dados de motilidades altas. Para vermes fêmeas, o ajuste de altas motilidades é razoável, mas o mesmo não ocorre para os vermes machos.

A figura 27 é a primeira a evidenciar que o modelo IID para o efeito de verme não é suficiente para ajustar adequadamente os dados experimentais de motilidade. As figuras 28 e 29 reforçam este resultado, mostrando inclusive que ele é independente da distribuição usada.

Pode-se observar por comparação dos gráficos 27-29 que, também independentemente da distribuição, o ajuste dos modelos IID+IID e SEM foi muito semelhante entre eles, sendo muito superiores ao modelo IID. Entretanto, o modelo SEM mostrou-se mais útil, na prática, por permitir realizar a inferência usando-se uma pequena fração do tempo gasto no uso do outro modelo.

Nota-se[30] que o modelo com distribuição binomial negativa (figura 27) tem um ajuste razoável para uma motilidade mediana, onde observa-se apenas um

[28] Um ajuste considerado ideal gera um gráfico com três retas de regressão quantil **horizontais** (uma para cada quartil), com coeficientes lineares iguais a 0,25, 0,5 e 0,75, respectivamente. Um desvio das retas horizontais sugere a presença de efeitos sistemáticos que não foram contemplados na modelagem sob diagnóstico.

[29] Por inspeção, notou-se que os resíduos mostravam tendências diferentes para valores extremos de motilidade quando comparados a valores médios. Para facilitar a discussão dos resultados, dividiu-se os dados de motilidade nas classes chamadas: baixa (abaixo do percentil 20%), mediana (entre os percentis 20% e 80%) e alta motilidade (acima de 80%).

[30] Considera-se daqui em diante somente os modelos IID+IID e SEM que mostraram ajuste superior ao modelo IID.

pequeno desvio da horizontal para a reta tracejada referente ao 2º quartil, e um desvio um pouco mais acentuado para o terceiro. Para motilidades baixas, grandes desvios em relação a horizontal são observados, maiores quanto menor o preditor linear (e, consequentemente, a motilidade). Motilidades altas seguem a tendência oposta, provavelmente como resultado da censura à direita ou algum outro efeito semelhante.

A distribuição binomial negativa inflada por zeros (figura 28) teve, como um avanço em relação à distribuição BN, a capacidade de reduzir o desvio da horizontal em motilidades baixas. Nota-se, graficamente, em especial para vermes machos, um bom ajuste para motilidades baixas e medianas. Isto ocorre com a estimativa de um grau (não-nulo) de inflação por zeros dado pelo parâmetro α (2,53 ±0,16 para os vermes machos, e 1,71 ±0,19 para vermes fêmeas). O principal resultado deste diagnóstico foi atestar a necessidade de aplicar um modelo para o excesso de zeros. Isto é interpretado como uma consequência, pelo menos parcialmente, da existência de um limite de quantificação para a motilidade (como antes caracterizado - figura 3).

Finalmente, a distribuição Poisson Generalizada parece ser adequada para motilidades medianas (figura 29; modelos IID+IID e SEM); para motilidades baixas, nota-se graficamente a heterocedasticidade anteriormente observada para a distribuição BN, mas um bom ajuste da medida de centralidade, já que o segundo quartil se mostra basicamente como uma reta horizontal. Para vermes machos, ainda se pode notar graficamente que a medida de centralidade também é boa para motilidades altas, enquanto que para fêmeas o desvio não é tão acentuado. Não foi estudada a versão inflada por zeros da distribuição GP, mas é esperado que ela seja adequada no ajuste para motilidades baixas, assim como foi alcançado para a distribuição BN inflada por zeros.

Como é sabido, a tarefa de uma análise estatística não se restringe à análise dos dados. A análise realizada tem o poder de propor melhorias no planejamento experimental. Com estes resultados pode-se sugerir que mudanças na configuração experimental poderiam resolver o problema no ajuste de valores altos de motilidade: possivelmente, o ajuste poderia ser muito melhor se o intervalo de tempo entre a captura de duas fotomicrografias fosse menor, pois resultaria em medidas consequentemente menores de motilidade. Elas provavelmente cobririam uma faixa de motilidade que tivesse um comportamento estritamente linear, como observado

na primeira porção da figura 18. Entretanto, uma consequência indesejada desta mudança no tempo entre a aquisição de duas bioimagens seria possivelmente o crescimento da autocorrelação entre as medidas consecutivas. Prevendo isso, pode-se tentar prevenir esta inconveniência alterando o padrão de captura de imagens[31].

Para uma comparação entre os modelos testados de melhor desempenho segundo a análise de resíduos, utilizou-se a medida WAIC. A tabela 3 mostra os resultados numéricos, com uso do modelo SEM, para a distribuição GP (com cada um dos modelos testados para o efeito de tratamento) e BN inflada por zeros. Estes resultados mostram que nenhum modelo testado para o efeito de composto *β* teve um grande destaque em qualidade para o conjunto de compostos estudados. Isto é um resultado da pouca similaridade entre os compostos avaliados, que não permitiu explorar as vantagens do modelo BYM. Apesar disso, este resultado mostra que o uso das redes não foi danoso à modelagem quando comparado ao modelo básico.

Por outro lado, a comparação entre as distribuições ZINB e GP mostra um aumento na qualidade de ajuste, com uma diferença absoluta [32] de 900, aproximadamente. Isto demonstra que a distribuição GP (com WAIC menor) é preferida a ZINB, mesmo na ausência de correção para excesso de zeros.

Tabela 3: Valores de WAIC (medida comparativa de qualidade ou do poder preditivo entre modelos ajustados)

Modelo p/ o efeito *β*	Vermes machos		Vermes fêmeas	
	GP	ZINB	GP	ZINB
básico	374130,52	-	232921,47	-
Tanimoto	374130,61	-	232921,81	-
Scaffolds	374130,46	375034,11	232921,61	233824,12

Como exemplo de aplicação dos resultados do modelo GP-*Scaffolds*-SEM, pode-se reconhecer compostos que apresentaram atividade contra os vermes adultos (figura 30). Para isso ser feito de forma automatizada, pode-se calcular a

[31] Uma alternativa seria capturar fotomicrografias sequencialmente com intervalos de tempo alternados, por exemplo 150 e 300 ms. Destas extrair dados somente dos pares de fotos espaçadas por 150 ms. Isto reduziria a dependência não modelada entre as medidas de motilidade.

[32] A diferença absoluta entre os valores de WAIC é uma medida adequada de comparação entre dois modelos, ao invés da razão entre eles. Isto ocorre porque os valores de WAIC são proporcionais ao logaritmo da função *likelihood* (GELMAN *et al*, 2014).

probabilidade *a posteriori* do efeito *β* ter sinal oposto ao valor estimado.[33] Ela pode ser usada para tentar produzir uma medida de evidência estatística do efeito estimado de um composto. Nota-se que o modelo foi capaz de caracterizar corretamente o grupo controle-positivo (isto é, tratado com praziquantel), cujo efeito de composto estimado foi negativo para todos os tempos estudados. Ou seja, recuperou-se o efeito (estatisticamente significativo) supressor da motilidade dos vermes esperado após tratamento com PZQ.

Foram construídos *scripts* em R, para futuras aplicações destes modelos, como parte integrante do *pipeline* de análise de ensaios fenotípicos de *S. mansoni* adulto.

Na figura 31, são apresentadas as estimativas do efeito de tratamento usando dados agregados por verme, separadamente para cada tempo estudado. São apresentados, para cada grupo de tratamento, as estimativas pontuais (média entre as replicatas) e intervalares (intervalo de confiança de 95%, assumindo distribuição *t-student*[34]). Para fazer uma comparação entre os resultados dos modelos propostos neste trabalho (ilustrados na figura 30) e os resultados oriundos da análise de dados agregados é preciso apresentá-los em uma mesma escala de medida. Para isso, a análise de dados agregados foi feita com uso da variável *logit* FNR. Comparativamente, é possível notar graficamente a diferença entre as abordagens e verificar a vantagem da modelagem empregada neste trabalho (e ilustrada pela figura 30). Nota-se, pela comparação das figuras 30 e 31, que a estimação dos intervalos de confiança com uso do modelo GP-*Scaffolds*-SEM foi mais robusta que a estimação usando dados agregados.

[33] Por definição, quanto menor ela é, maior é a evidência, estatisticamente, de que o efeito de tratamento é diferente do grupo controle; o seu valor é esperado ser próximo de 0,5 quando o efeito estimado é igual a 0.

[34] Quando se supõe que os dados sobre uma variável estudada compõem uma amostra aleatoria simples, de tamanho n, obtida de uma população com distribuição normal com média e variância desconhecidas, é possível estimar um intervalo de confiança calculando-se quantis da distribuição *t* de *Student* com n – 1 graus de liberdade (BUSSAB e MORETTIN, 2017).

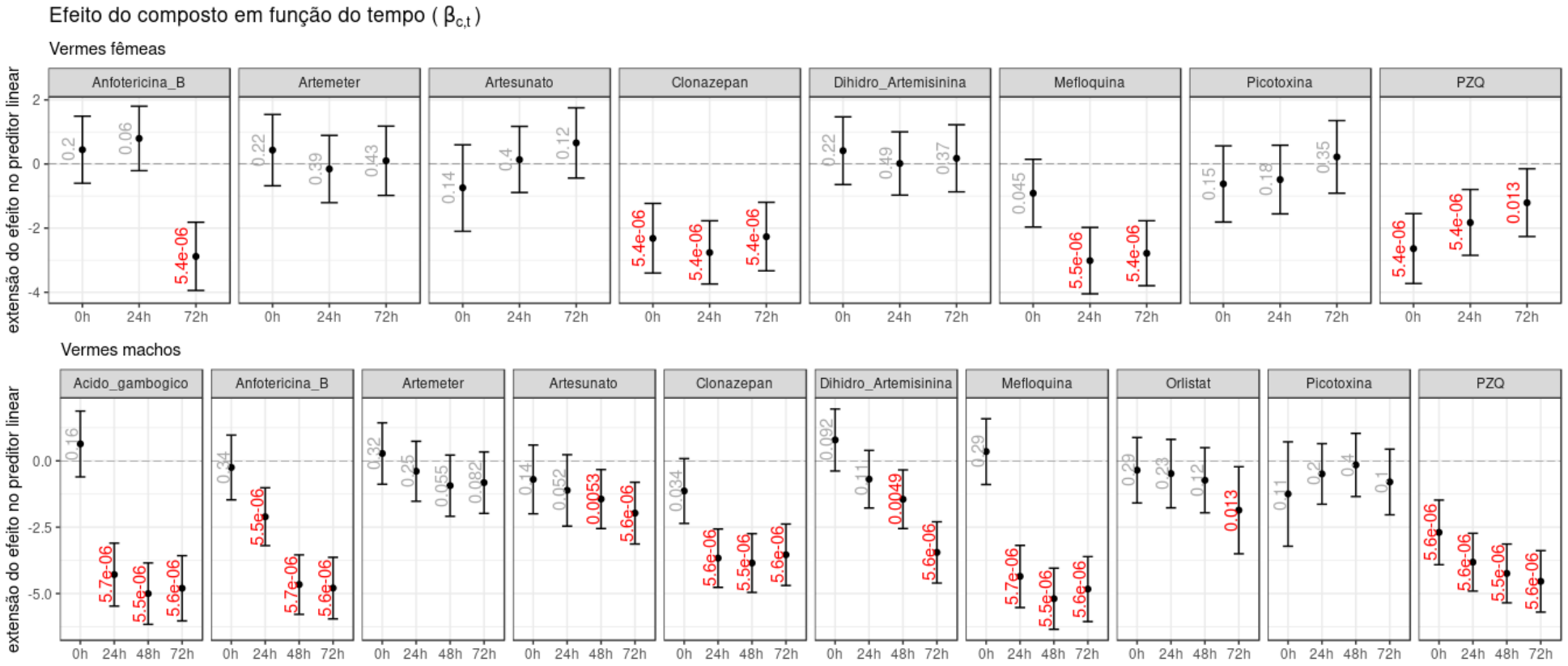


Figura 30: Resultados dos efeitos dos compostos (relativo ao grupo controle), aplicando-se a distribuição GP e os modelos *Scaffolds* e SEM. Os intervalos de confiança de 95% ao redor do valor estimado (média) são representados como barras verticais. Os números ao lado das barras indicam a probabilidade a *posteriori* do efeito *β* ter sinal oposto ao valor estimado. Em vermelho, destaque para os efeitos considerados estatisticamente diferentes do efeito do grupo controle e, em cinza, os efeitos não significativos estatisticamente (usando-se um valor de *cut-off* de 0,025).

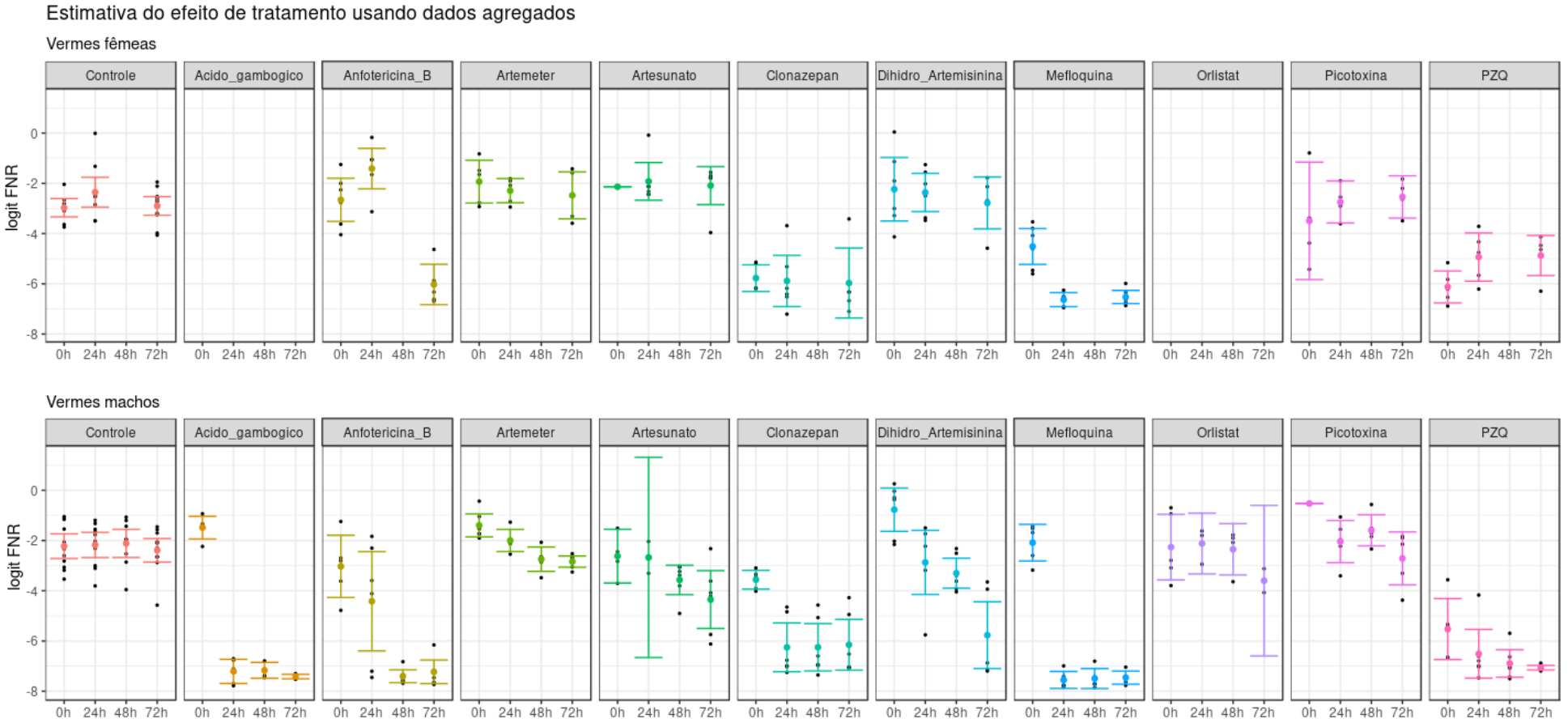


Figura 31: Estimativa do efeito de tratamento usando dados agregados por verme. Os círculos pretos indicam o valor médio das medidas de *logit* FNR de cada verme e tempo. As barras verticais representam estimativas do intervalo de confiança de 95% para o valor médio de *logit* FNR, com a suposição de que os resíduos tem distribuição *t-student*, com ($n_{c,t}$ – 1) graus de liberdade, onde $n_{c,t}$ é o número de replicatas tratamento × tempo. Os círculos coloridos (contidos nas barras verticais) representam as estimativas pontuais dos efeitos de tratamento médios para cada tratamento/composto ao longo do tempo de ensaios. Quando a estimativa do intervalo de confiança não foi possível, por $n_{c,t}$ = 1, a barra vertical foi suprimida do gráfico.

4.2.2. Modelos para experimentos do tipo Dose-Resposta

Neste tipo de experimento, buscou-se desenvolver uma adaptação para experimentos do tipo Dose-Resposta, a partir do modelo produzido para a triagem de compostos. Para isso, substituiu-se os modelos para efeito de tratamento explorados na seção anterior por efeitos não lineares que levam em conta a dose do composto, como descrito anteriormente.

Na seção anterior concluiu-se que os modelos IID+IID e SEM são basicamente equivalentes entre si, mas o modelo SEM foi mais eficiente computacionalmente. Além disso, a distribuição GP foi considerada a mais adequada para o ajuste dos dados experimentais de motilidade. Por isso, somente foram testados os modelos com esta distribuição e com uso do modelo SEM para efeito a nível de verme.

Os gráficos da figura 32 ilustram as curvas Dose-Resposta log-logística generalizada do composto oltiplaz para vermes fêmeas e machos, respectivamente. Da mesma forma, os gráficos da figura 33 apresentam as curvas Weibull estimadas. Nestas figuras[35], as curvas Dose-Resposta são dadas pelo efeito do tratamento β_t, analogamente aos efeitos estimados e apresentados na figura 30. As diferenças são somente a escala que se apresenta o efeito - pois a curva Dose-Resposta é dada por $\exp(\beta_t)$ em função da dose - e o fato de o efeito ser uma função (da dose). A tabela 4 completa a informação dos gráficos com as medidas-resumo das distribuições *a posteriori* dos parâmetros de interesse das curvas Dose-Resposta (d, b e EC_{50}). A comparação dos resultados numéricos mostra que: as estimativas do parâmetro *d* (o limite superior da curva) para cada tempo têm valores sistematicamente menores quando o modelo Weibull é usado, mas são bem próximos aos resultados obtidos com o modelo log-logístico generalizado; os resultados para o parâmetro b_t também não foram muito diferentes entre um modelo e outro, com única exceção para o caso de vermes fêmeas, 24h, onde o modelo Weibull mostrou-se com dificuldade de ajustar os dados aparentemente; e os resultados de EC_{50} indicam que o modelo Weibull têm a tendência de estimar um

[35] Apesar da medida de motilidade modelada ser a variável FN, como definida anteriormente, as curvas Dose-Resposta obtidas podem ser melhor comparadas aos dados experimentais na escala de motilidade relativa / deslocamento relativo (correspondente à variável FNR). Além disso, mesmo que a modelagem tenha sido feita com dados a nível de medida, este são omitidos e os gráficos mostram somente os dados pontuais de motilidade a nível de verme.

conjunto de valores de EC_{50} nem muito pequenos nem muito elevados, enquanto o modelo log-logístico não apresenta esta característica. Para a determinação de qual é o modelo mais adequado para este caso de estudo em especial, seriam necessárias mais análises comparativas[36]. Em aplicações destes modelos para outros dados, razões teóricas também podem ser decisivas para a escolha de qual modelo se adotar.

Os efeitos a nível de verme são responsáveis por modelar a dispersão dos dados pontuais apresentados nas figuras. Como eles são construídos como efeitos gaussianos *no preditor linear*, eles não apresentam distribuição gaussiana (nem simétrica) na escala de motilidade relativa (FNR), devido ao efeito da função-*link*. Isto permite levar em conta a distribuição natural assimétrica dos dados observada na análise exploratória. Dessa forma, vermes que seriam considerados "*outliers*" em uma modelagem padrão, e impactariam negativamente na estimação dos parâmetros de interesse, não parecem distorcer os resultados ilustrados pelas figuras. Entretanto, é sugestivo a partir dos gráficos que algum efeito do tipo hormesis, não contemplado pelos modelos usados, pode estar presente. Note, em especial, a motilidade de vermes fêmeas tratadas com doses moderadas de oltipraz. Elas mostram uma tendência de crescimento inicial da motilidade com o aumento da dose, em relação ao grupo controle, que não pode ser capturada por estes modelos.

A figura 34 mostra, como um exemplo de aplicação, as curvas Dose-Resposta obtidas com uso do modelo Weibull com efeito hormesis do tipo Cedergreen-Ritz-Streibig incluso. Nota-se que, com a adição de um efeito de hormesis na construção do efeito de tratamento β_t, a incerteza associada a ele cresce, refletindo a necessidade de se estimar mais um parâmetro com o mesmo conjunto de dados de antes. Apesar disso, o ajuste aos dados pontuais (principalmente, após 72h de tratamento) foi visualmente melhor quando comparado aos resultados da figura 33. O problema com a modelagem do efeito hormesis testada aqui é a recomendação de se fixar *a priori* os parâmetros que governam a taxa de crescimento do efeito hormesis, antes da estimação propriamente dita dos parâmetros de interesse (CEDERGREEN et al, 2005). A escolha dos parâmetros adotados vai depender de uma avaliação prévia dos dados pelo futuro usuário do *script,* o que torna o procedimento não mais automatizado nesta etapa, a princípio.

[36] Outras análises não foram feitas, pois o escopo desta etapa do trabalho foi somente a construção de diferentes opções de modelos para experimentos do tipo Dose-Resposta.

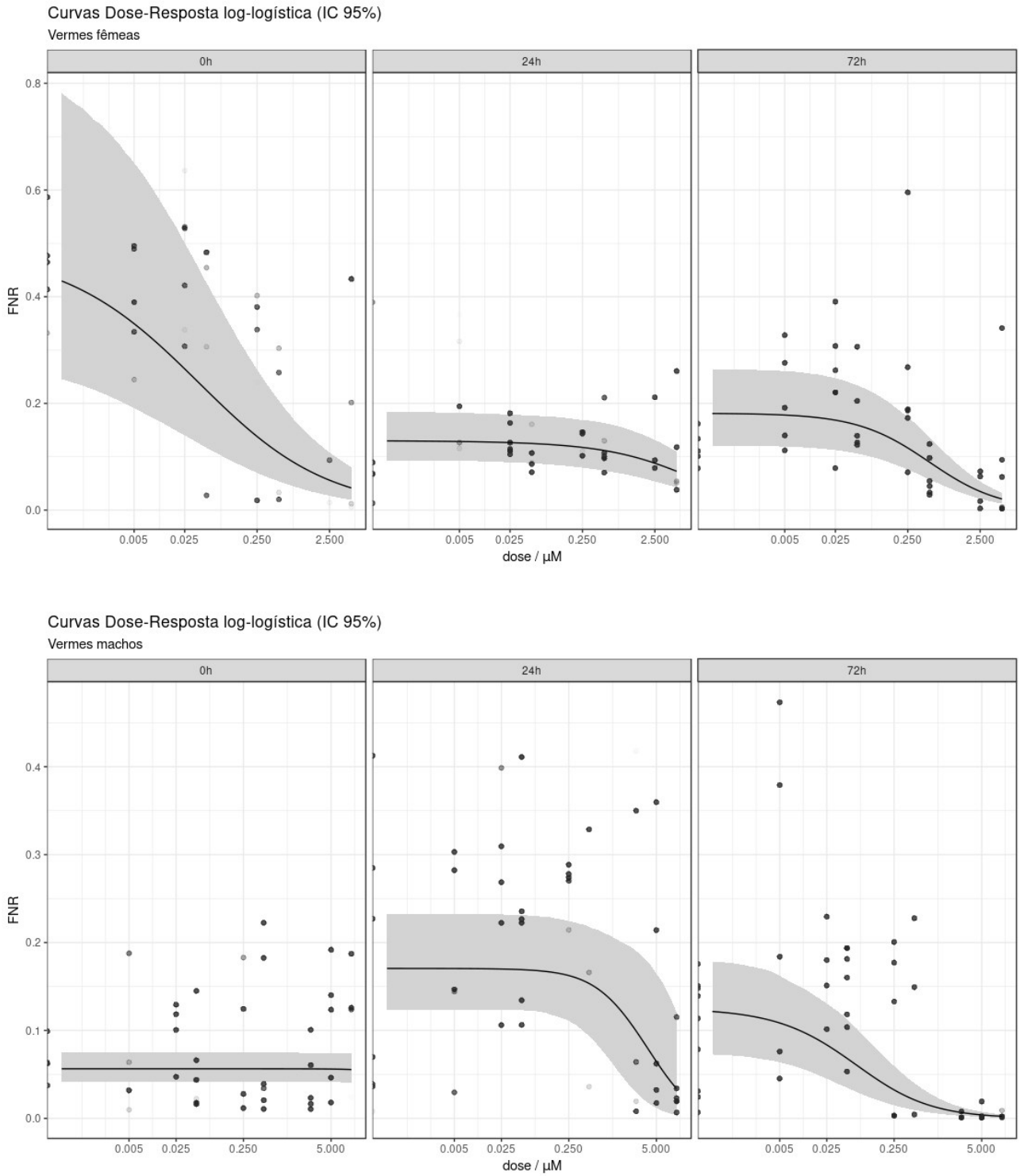


Figura 32: Curvas Dose-Resposta obtidas usando o modelo log-logístico generalizado de quatro parâmetros. O parâmetro limite inferior foi fixado em zero. Os dados pontuais representam a média de motilidade experimental a nível de verme, em termos de FNR (onde dados pontuais do grupo controle são apresentados na extremidade esquerda de cada gráfico). As medidas-resumo, em função da dose, da distribuição *a posteriori* de $\exp(\beta_t)$ são dadas pela curva preta (média) e pela área cinza (IC 95%). A abscissa é apresentada na escala logarítmica.

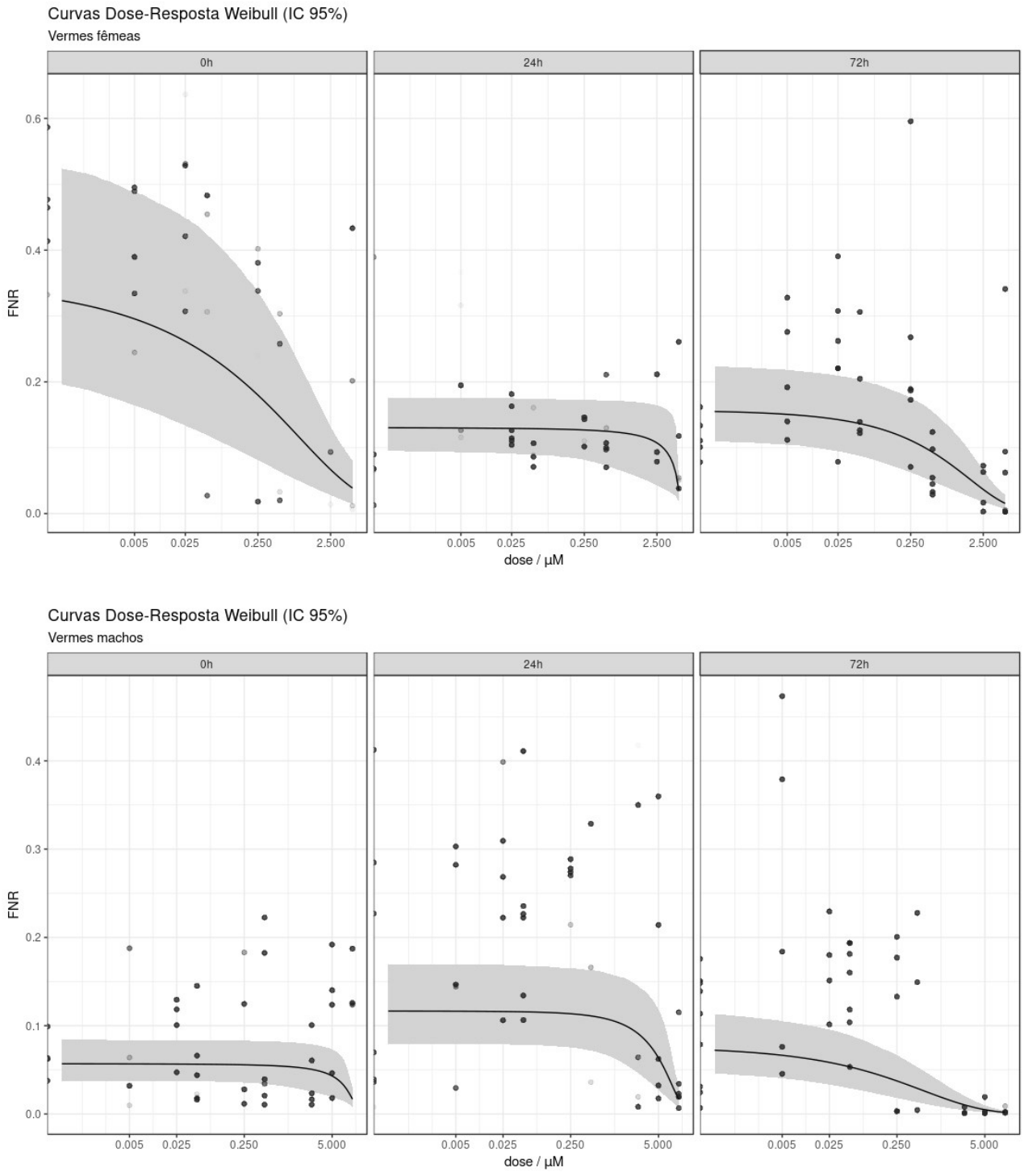


Figura 33: Curvas Dose-Resposta obtidas usando o modelo Weibull I de quatro parâmetros. O parâmetro limite inferior foi fixado em zero. Os dados pontuais representam a média de motilidade experimental a nível de verme, em termos de FNR (onde dados pontuais do grupo controle são apresentados na extremidade esquerda de cada gráfico). As medidas-resumo, em função da dose, da distribuição *a posteriori* de $\exp(\beta_t)$ são dadas pela curva preta (média) e pela área cinza (IC 95%). A abscissa é apresentada na escala logarítmica.

Tabela 4: Medidas-resumo dos *parâmetros* das curvas Dose-Resposta

	Vermes fêmeas						Vermes machos					
	Log-logístico			Weibull			Log-logístico			Weibull		
	m	SD	IC_{95}	m	SD	IC_{95}	m	SD	IC_{95}	m	SD	IC_{95}
d_{0h}	0,48	0,15	0,29 0,89	0,35	0,09	0,22 0,56	0,06	0,01	0,04 0,07	0,06	0,01	0,04 0,08
d_{24h}	0,13	0,02	0,09 0,18	0,13	0,02	0,10 0,18	0,17	0,03	0,12 0,18	0,12	0,02	0,08 0,17
d_{72h}	0,18	0,04	0,12 0,26	0,16	0,03	0,11 0,23	0,13	0,03	0,07 0,18	0,08	0,02	0,05 0,12
b_{0h}	0,49	0,06	0,39 0,62	0,41	0,14	0,21 0,75	2,72	1,08	1,15 5,22	3,03	2,62	0,42 9,96
b_{24h}	0,75	0,23	0,38 1,27	7,48	12,4	0,42 37,8	1,67	0,47	0,95 2,74	1,87	1,11	0,64 4,78
b_{72h}	0,88	0,08	0,76 1,06	0,68	0,20	0,38 1,14	0,80	0,10	0,63 1,00	0,46	0,09	0,31 0,66
$EC50_{0h}$	0,04	0,03	0,01 0,11	0,35	0,32	0,01 1,20	85,9	18,2	82,9 125,4	7,45	1,81	2,75 9,64
$EC50_{24h}$	7,43	3,26	2,10 14,3	3,92	1,01	1,04 4,93	4,25	3,74	0,78 14,69	5,17	1,62	1,91 8,09
$EC50_{72h}$	0,49	0,11	0,31 0,73	0,83	0,44	0,17 1,82	0,07	0,04	0,02 0,19	0,28	0,20	0,04 0,78

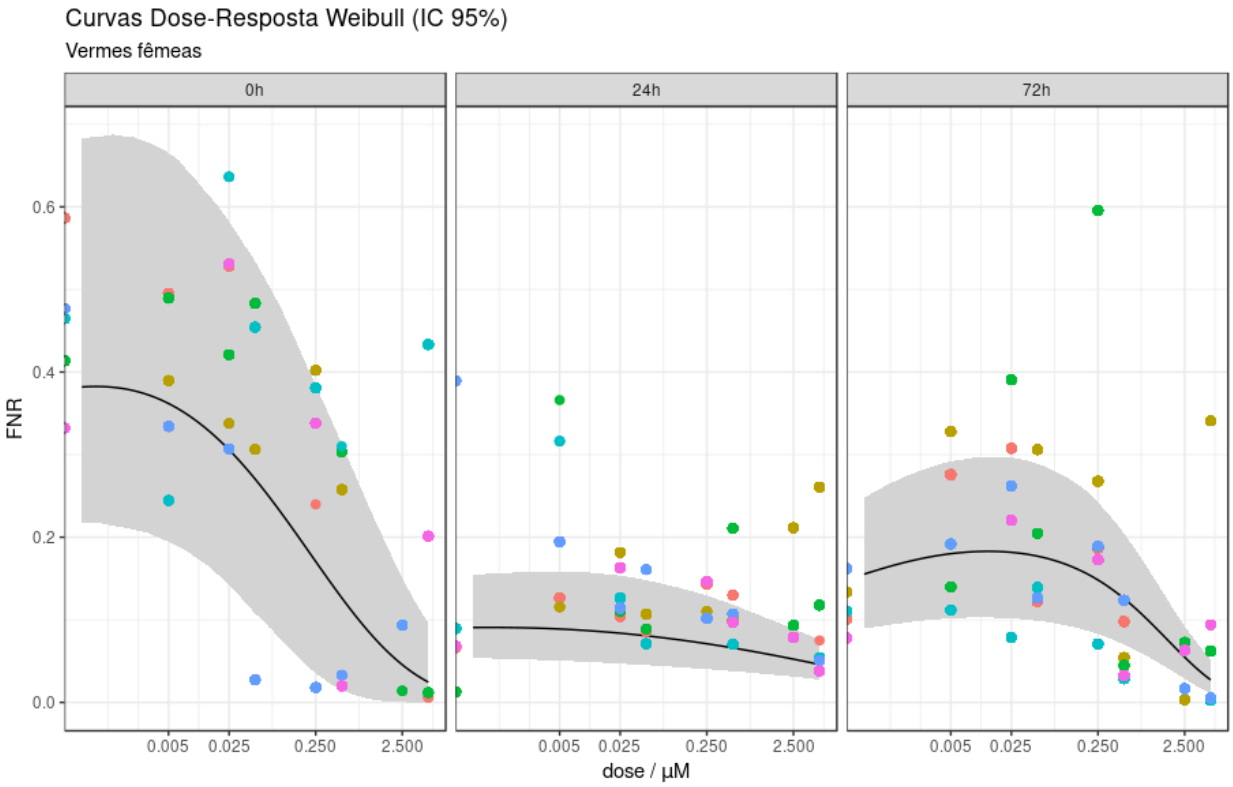


Figura 34: Curvas Dose-Resposta obtidas usando o modelo Weibull com efeito hormesis do tipo Cedergreen-Ritz-Streibig modificado de cinco parâmetros, para vermes fêmeas. O parâmetro limite inferior foi fixado em zero. Os dados pontuais representam a média de motilidade experimental a nível de verme, em termos de FNR (onde dados pontuais do grupo controle são apresentados na extremidade esquerda de cada gráfico). As medidas-resumo, em função da dose, da distribuição a posteriori de exp(βt) são dadas pela curva preta (média) e pela área cinza (IC 95%). A abscissa é apresentada na escala logarítmica. As cores diferentes dos dados pontuais servem para identificar os vermes de um grupo de tratamento em tempos diferentes.

4.3 Análise Multivariada

Com o intuito de testar a viabilidade de um modelo preditivo multivariado, construído a partir de métodos já consagrados de análise multivariada - e que mantenha características semelhantes àquelas do modelo univariado usado na primeira etapa do trabalho - escolheu-se seguir as etapas abaixo:

a) Análise exploratória dos dados pré-tratamento e transformação das variáveis;

b) Redução da dimensionalidade por PCA dos dados de pré-tratamento transformados (sugerida pelos resultados da etapa *a*) e geração das funções *score* associadas a cada um dos componentes principais preservados;

c) Agrupamento *fuzzy*[37] dos dados de tratamento (a fim de identificar diferentes fenótipos e distingui-los do fenótipo apresentado pelo grupo controle), em termos dos componentes principais e suas funções *score* associadas, e determinação do *cluster* Controle;

d) Construção de modelos do tipo árvore de regressão com efeitos aleatórios estimados a nível de verme[38], tomando-se como variável-resposta o grau de pertencimento ao *cluster* Controle (obtido na etapa c).

Diferentemente da seção anterior - onde se produziu modelos para análises univariadas com vantagens na interpretação dos parâmetros estimados - o foco da análise multivariada deste trabalho é a busca de modelos com bom desempenho *preditivo.* Por isso, esta parte do trabalho tem como característica uma abordagem mais pragmática, onde uma análise detalhada de cada uma das variáveis não se fez interessante.

4.3.1. **Análise Exploratória**

Com um total de 92 variáveis[39] originalmente - divididas em quatro grupos correspondentes aos módulos de medida *MeasureGranularity*, *MeasureTexture*,

[37] É interessante esclarecer a escolha de aplicação da técnica de agrupamento fuzzy. Neste trabalho, empregou-se este tipo de agrupamento a fim de gerar uma variável contínua (o grau de pertencimento ao grupo controle) que pudesse ser usada na etapa d) construção de um modelo do tipo árvore de regressão com efeitos aleatórios.

[38] Efeito a nível de verme, que foi anteriormente mostrado ser importante na modelagem da motilidade, é esperado ser importante também na análise multivariada, por extensão.

[39] O *pipeline* de processamento de imagens em *software* CellProfiler (cujo *script* é apresentado no Anexo 1) gera rótulos descritivos (e muito longos) para cada variável. Por isso, na apresentação dos resultados, substituiu-se os rótulos originais por outros que informam somente os módulos usados para obter as medidas e índices numéricos. Como não se tem o objetivo, neste trabalho, de fazer uma análise detalhada de todas as

CalculateImageOverlap e *MeasureImageIntensity* - inicia-se uma análise exploratória, a fim de se estudar a correlação entre as variáveis e suas distribuições empíricas dos dados de pré-tratamento. O primeiro resultado da análise foi a identificação de 3 variáveis do módulo *CalculateImageOverlap* (O_009, O_011 e O_014) que não tinham utilidade e, por isso, foram descartadas[40].

Um estudo geral da correlação entre as diferentes variáveis foi feito para se obter informações sobre o grau de dependência entre elas e avaliar a forma de se representar os dados mais concisamente. Gráficos da matriz de correlação entre as variáveis são apresentados nas figuras 35-38. Em geral, pode-se observar alto grau de correlação linear entre variáveis de um mesmo módulo (cuja maioria dos coeficientes são acima de 0,9), com exceção somente das variáveis do módulo *MeasureGranularity*. Em contraste, variáveis de um módulo de medida tendem a apresentar baixíssima correlação com variáveis de um outro módulo de medida. Como uma exceção nota-se o par *CalculateImageOverlap-MeasureImageIntensity,* no caso de vermes machos, onde os valores absolutos dos coeficientes de correlação são baixos a moderados. A outra exceção é o par *MeasureImageIntensity-MeasureTexture,* com as mesmas características.

Resulta-se desta análise que a informação contida neste conjunto de variáveis estudadas é altamente redundante, mas cada módulo de medida é pouco correlacionado com os demais. Apesar dos módulos *MeasureTexture e MeasureGranularity* terem finalidades semelhantes (ambos buscam avaliar a textura das imagens), a correlação entre suas variáveis não foi grande, em contraste do que se poderia esperar inicialmente.

A alta correlação entre as variáveis sugere o uso de um método que permita a redução da dimensionalidade dos dados. Para isso, escolheu-se realizar a Análise de Componentes Principais (PCA) e preservar um número reduzido destes para representar os dados. Como a PCA geralmente produz melhores resultados quando as variáveis são relativamente simétricas, seguiu-se com a avaliação das distribuições univariadas empíricas das variáveis originais dos dados de pré-tratamento.

---

variáveis, isto foi suficiente. Entretanto, uma relação entre os rótulos originais e os substitutos das variáveis são apresentados no Anexo 3.

[40] As “variáveis” na verdade eram **constantes** (assumindo os valores 0, 1 e 1, respectivamente): logo, não forneciam informação alguma, e estavam presentes como simples artefatos do processamento das imagens.

Como se observa na figura 39, as distribuições das variáveis originais do módulo *CalculateImageOverlap* são bastante assimétricas. Isto se confirma também para os módulos restantes e é independente do sexo do verme[41]. Devido a isto, foram testadas algumas transformações monótonas para tentar reduzir a assimetria observada. Em muitos casos, as transformações *log* ou *logit* não foram adequadas pela presença de valores nulos ou negativos nos dados originais, que estão fora do domínio destas funções. A aplicação da função raiz cúbica não apresenta esta desvantagem e foi a mais efetiva para esta finalidade e foi aplicada para todas as variáveis assimétricas positivas, enquanto as assimétricas para a esquerda[42] sofreram a transformação $T(X) = \sqrt[3]{(1 - X)}$. Como resultado final, os dados de pré-tratamento puderam ser representados por distribuições de maior simetria, ou, em geral, como uma mistura de distribuições essencialmente simétricas, como desejado.

[41] Os histogramas das variáveis dos outros módulos, para ambos os sexos, se encontram no anexo 4.

[42] Todas as variáveis assimétricas para esquerda estavam limitadas em módulo a valores entre 0 e 1 (como, por exemplo, a variável O_007 ilustrada na figura 39). Por isso, a transformação linear T'(X) = (1 - X) foi suficiente para inverter o sentido da sua assimetria. As variáveis originalmente assimétricas para a esquerda foram: O_007, O_010, O_012, O_013, O_015, I_019, I_021,I_023,I_025, I_026, I_032, T_049, T_050, T_051, T_052, T_057, T_058, T_059, T_060, T_082, T_083, T_084, T_085.

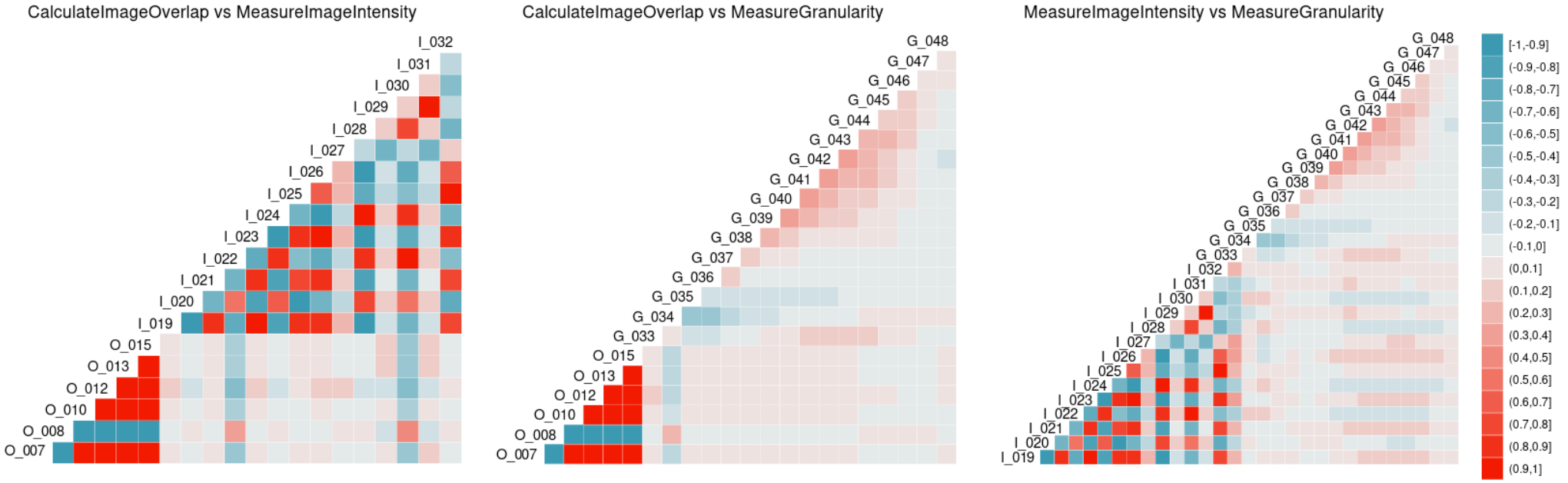


Figura 35: Matrizes de correlação ilustradas, comparando variáveis dos módulos *CalculateImageOverlap, MeasureImageIntensity* e *MeasureGranularity* para dados de vermes fêmeas.

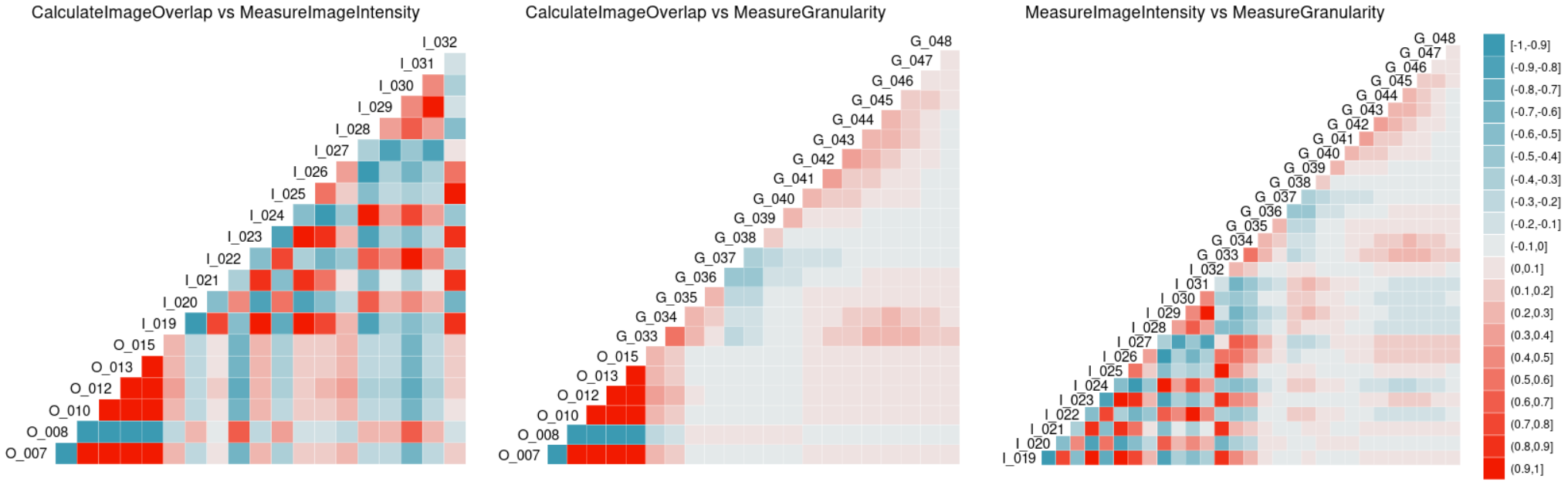


Figura 36: Matrizes de correlação ilustradas, comparando variáveis dos módulos *CalculateImageOverlap, MeasureImageIntensity* e *MeasureGranularity* para dados de vermes machos.

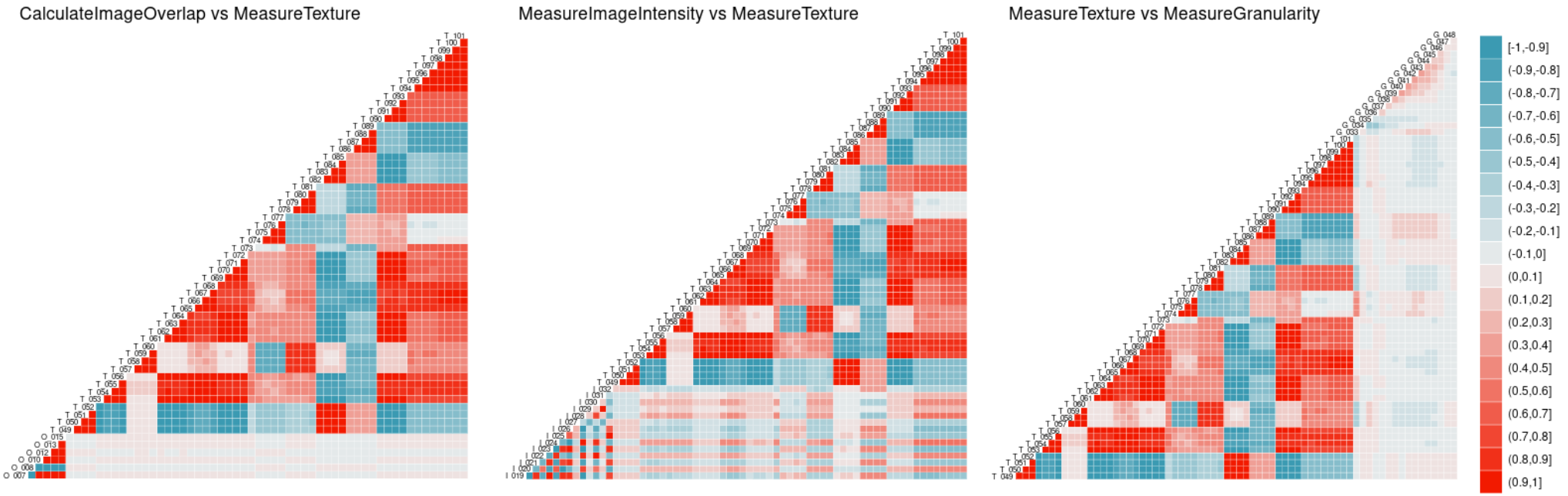


Figura 37: Matrizes de correlação ilustradas, comparando as variáveis do módulo *MeasureTexture* com as variáveis dos outros módulos restantes, para dados de vermes fêmeas.

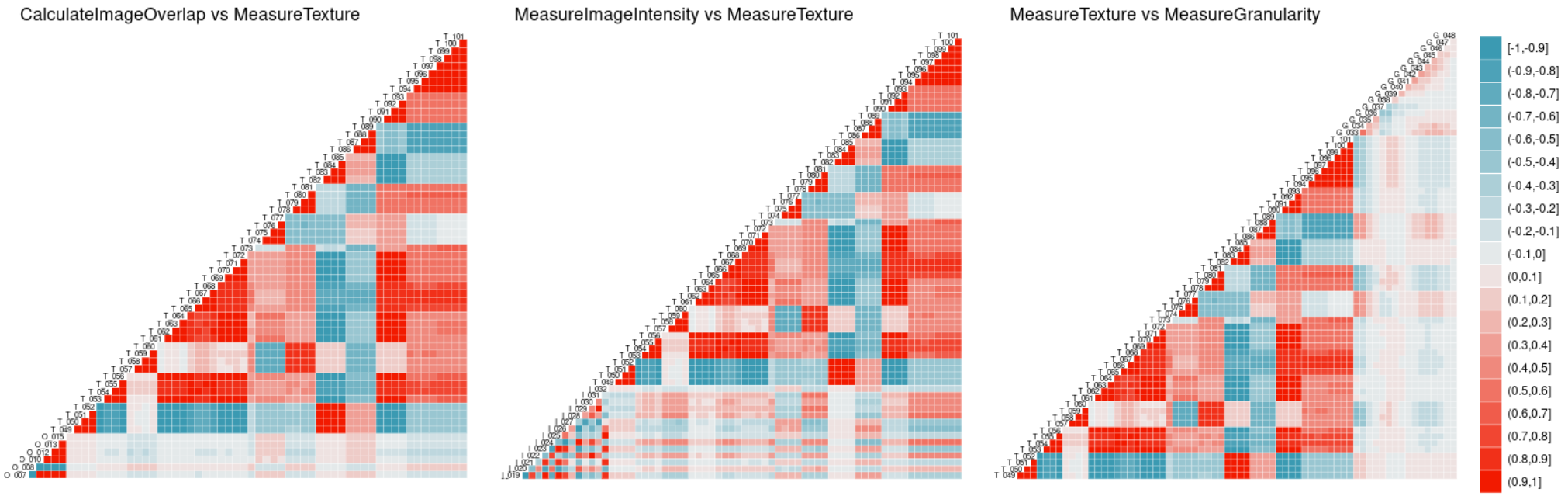


Figura 38: Matrizes de correlação ilustradas, comparando as variáveis do módulo *MeasureTexture* com as variáveis dos outros módulos restantes, para dados de vermes machos.

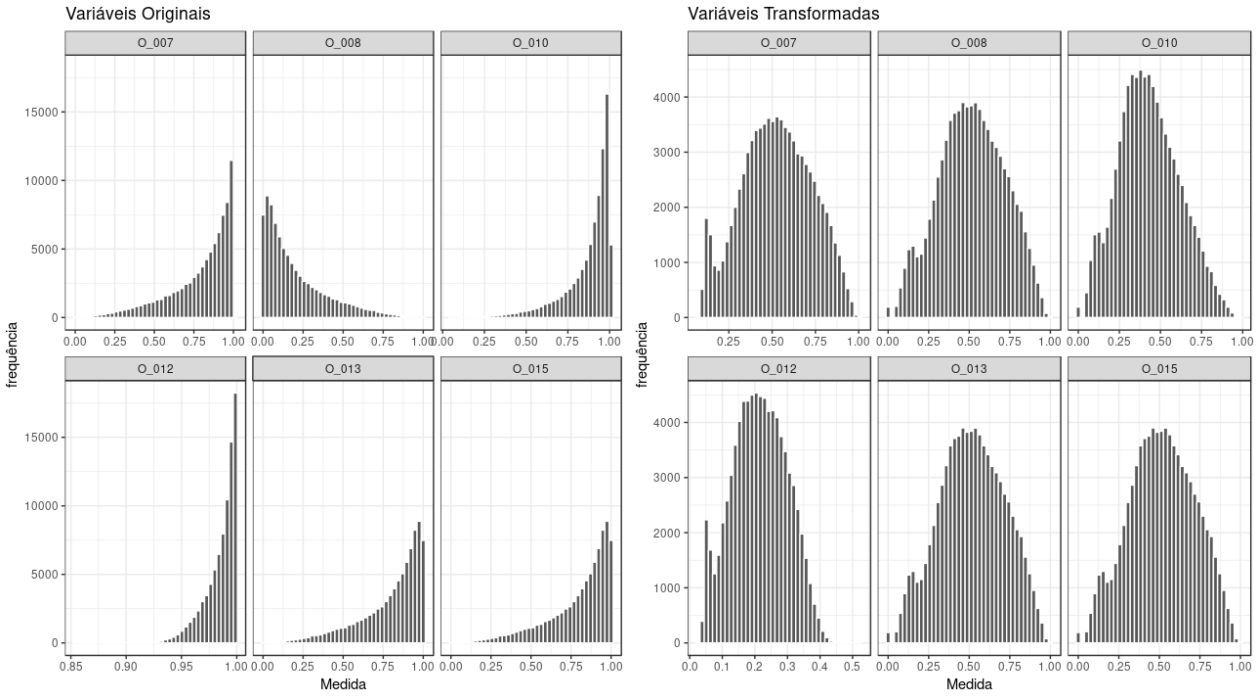


Figura 39: Comparação entre os histogramas das variáveis originais e transformadas do módulo *CalculateImageOverlap*, para dados de pré-tratamento em vermes fêmeas.

4.3.2. **Redução de Dimensionalidade**

O conjunto de dados de pré-tratamento original foi transformado como descrito anteriormente, e em seguida foi feita a identificação dos componentes principais para cada sexo separadamente (a figura 40 mostra os gráficos do autovalor associado a cada componente principal identificado). Utilizando-se o critério de Kaiser, dez componentes principais foram retidos tanto para fêmeas quanto para machos. Os 10 primeiros componentes principais respondem por 87,3% da variância dos dados de vermes fêmeas e 88,0% desta no caso dos vermes machos.

Analisando as cargas de cada componente principal retido, observa-se que o primeiro componente principal é essencialmente composto por contribuições das variáveis geradas pelos módulos de medida *MeasureImageIntensity* e *MeasureTexture* - como mostra a figura 41 para o caso de vermes fêmeas[43] - com predominância do segundo módulo. O segundo componente também tem grande

[43] Os gráficos dos componentes principais que não foram apresentados nos resultados se encontram no anexo 5.

participação das variáveis dos módulos *MeasureImageIntensity* e *MeasureTexture,* mas, neste caso, com predominância do primeiro. O terceiro componente principal é composto por contribuições importantes das variáveis do módulo *CalculateImageOverlap*, seguido das variáveis do módulo *MeasureImageIntensity,* e de 12 das 53 variáveis do módulo *MeasureTexture* (figura 42). O quarto componente tem um perfil semelhante ao anterior. Do quinto ao décimo componente há uma crescente contribuição das variáveis do módulo *MeasureGranularity,* que até então apresentaram cargas essencialmente nulas, concomitante a redução das cargas das variáveis dos outros módulos (figuras 43 e 44).

Estes resultados são coerentes, em geral, com os resultados da análise de correlação entre variáveis, pois as variáveis de um mesmo módulo contribuem juntas em cada componente principal, e com valor absoluto de suas cargas muito similar. No caso do módulo *MeasureGranularity*, a PCA identificou pelo menos 5 componentes diferentes gerados pelo espectro de medidas que caracteriza este módulo.

Com estes componentes principais e as cargas de cada variável, se pôde construir as funções *score*. A aplicação destas aos dados de tratamento foi feita para reduzir a sua dimensionalidade.

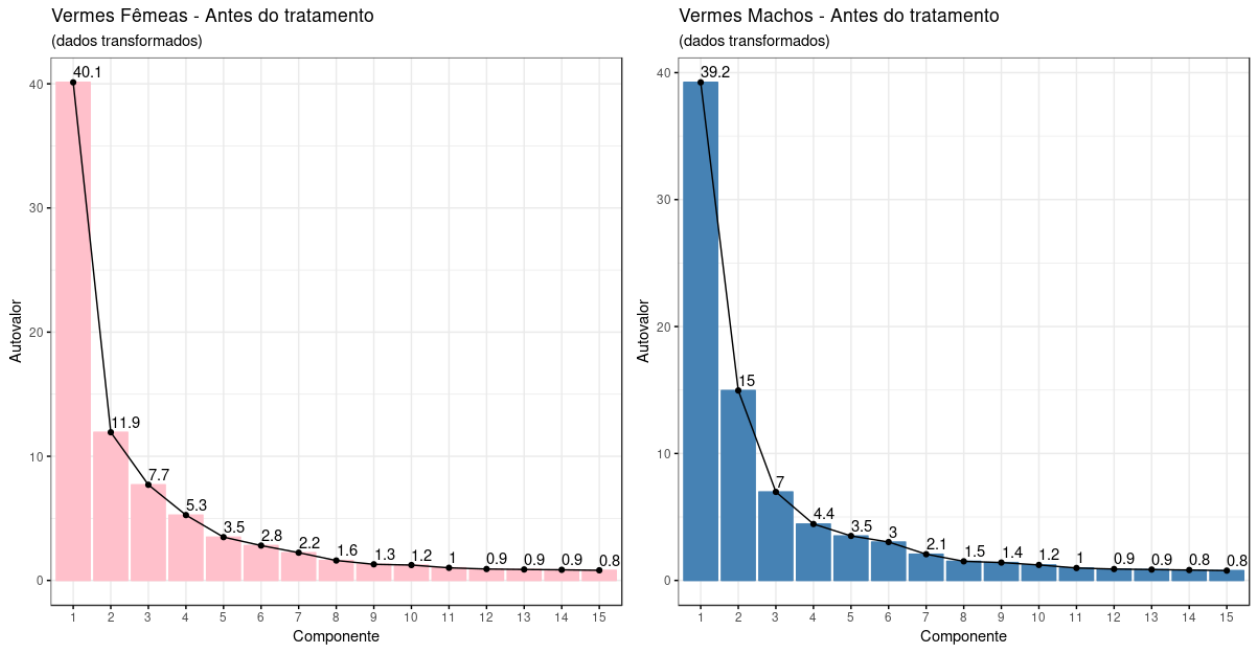


Figura 40: Gráficos dos autovalores para cada um dos 15 primeiros componentes principais para os resultados da PCA dos dados pré-tratamento, para cada sexo separadamente.

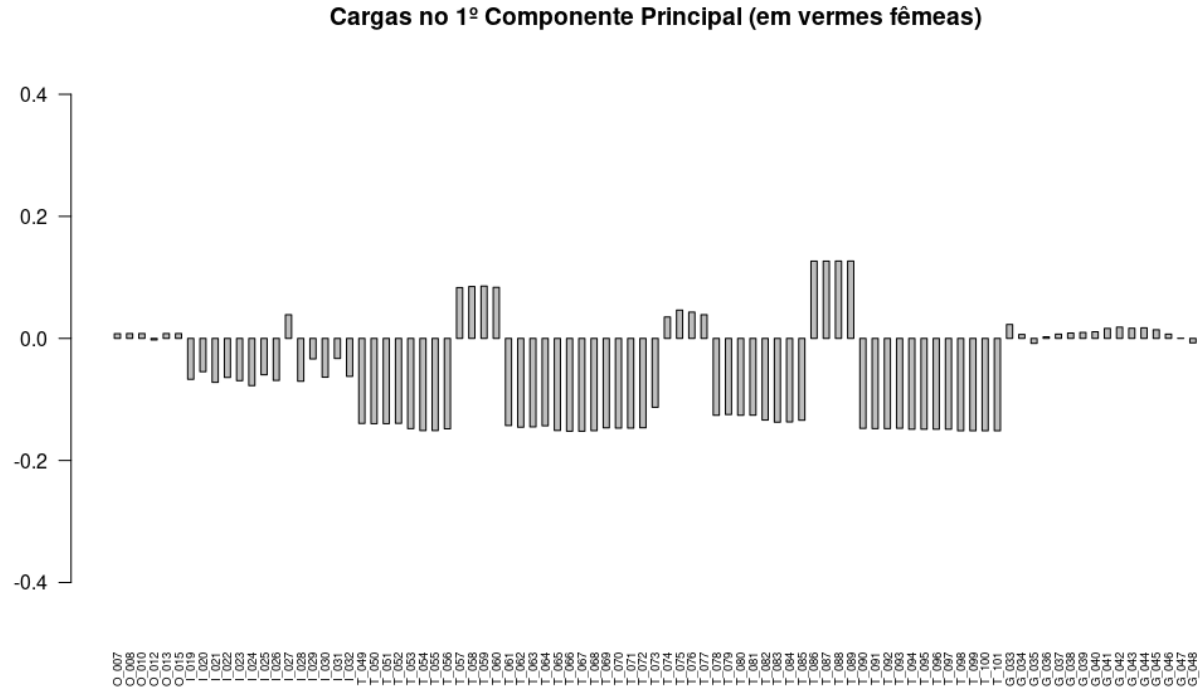


Figura 41: Gráfico de barras das cargas no 1º componente principal dos dados de pré-tratamento de vermes fêmeas.

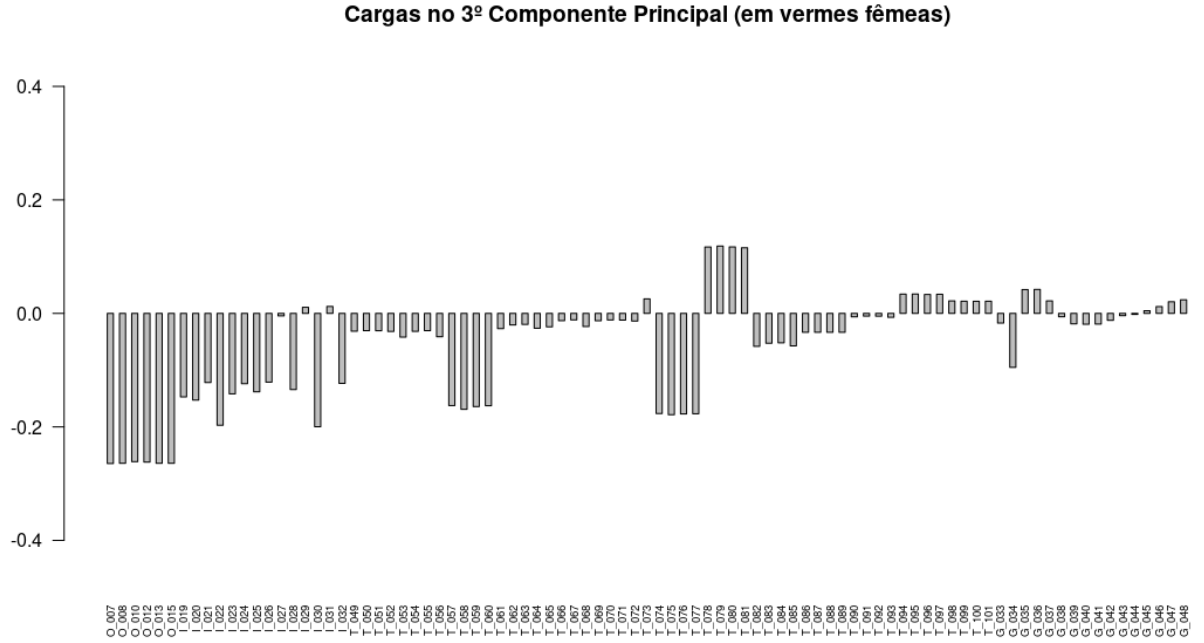


Figura 42: Gráfico de barras das cargas no 3º componente principal dos dados de pré-tratamento de vermes fêmeas.

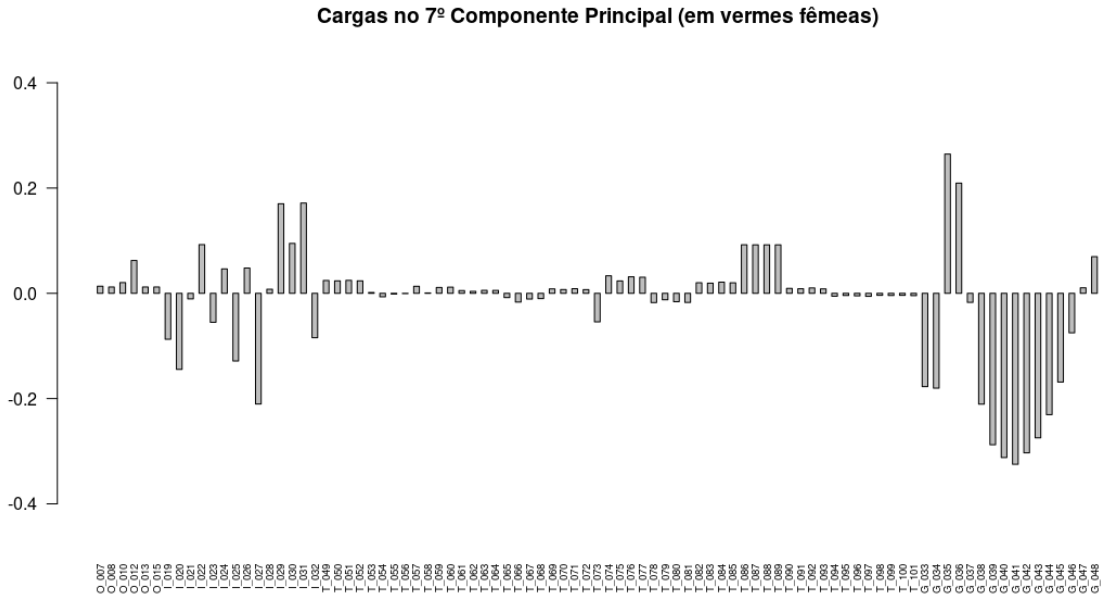


Figura 43: Gráfico de barras das cargas no 7º componente principal dos dados de pré-tratamento de vermes fêmeas.

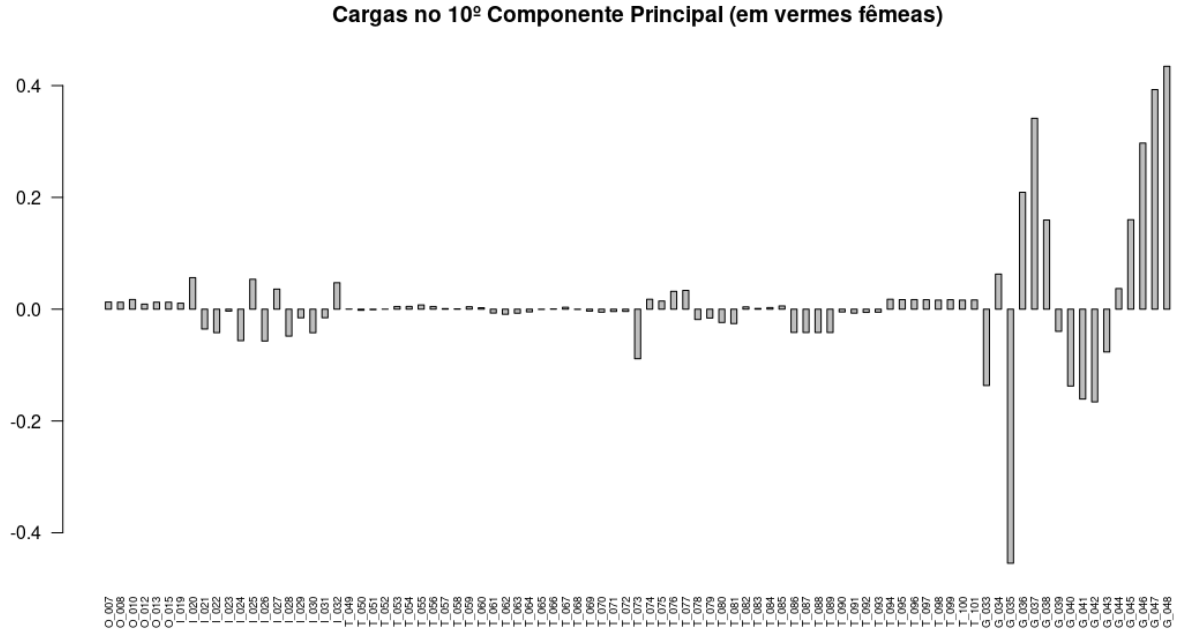


Figura 44: Gráfico de barras das cargas no 10º componente principal dos dados de pré-tratamento de vermes fêmeas**.**

4.3.3. **Resultados do agrupamento e da modelagem com Árvore de Regressão do grau de pertencimento ao *cluster* Controle**

Aplicando as funções *score* aos dados de tratamento para cada sexo separadamente, pôde-se obter os dados reduzidos contendo apenas 10 variáveis independentes, além dos metadados. Como descrito anteriormente, usou-se o algoritmo FANNY para detectar grupos e estimar graus de pertencimento de cada observação aos diferentes grupos detectados. As figuras 45 e 46 resumem os resultados das 20 replicatas obtidas em *boxplots* para cada configuração de parâmetros testada.

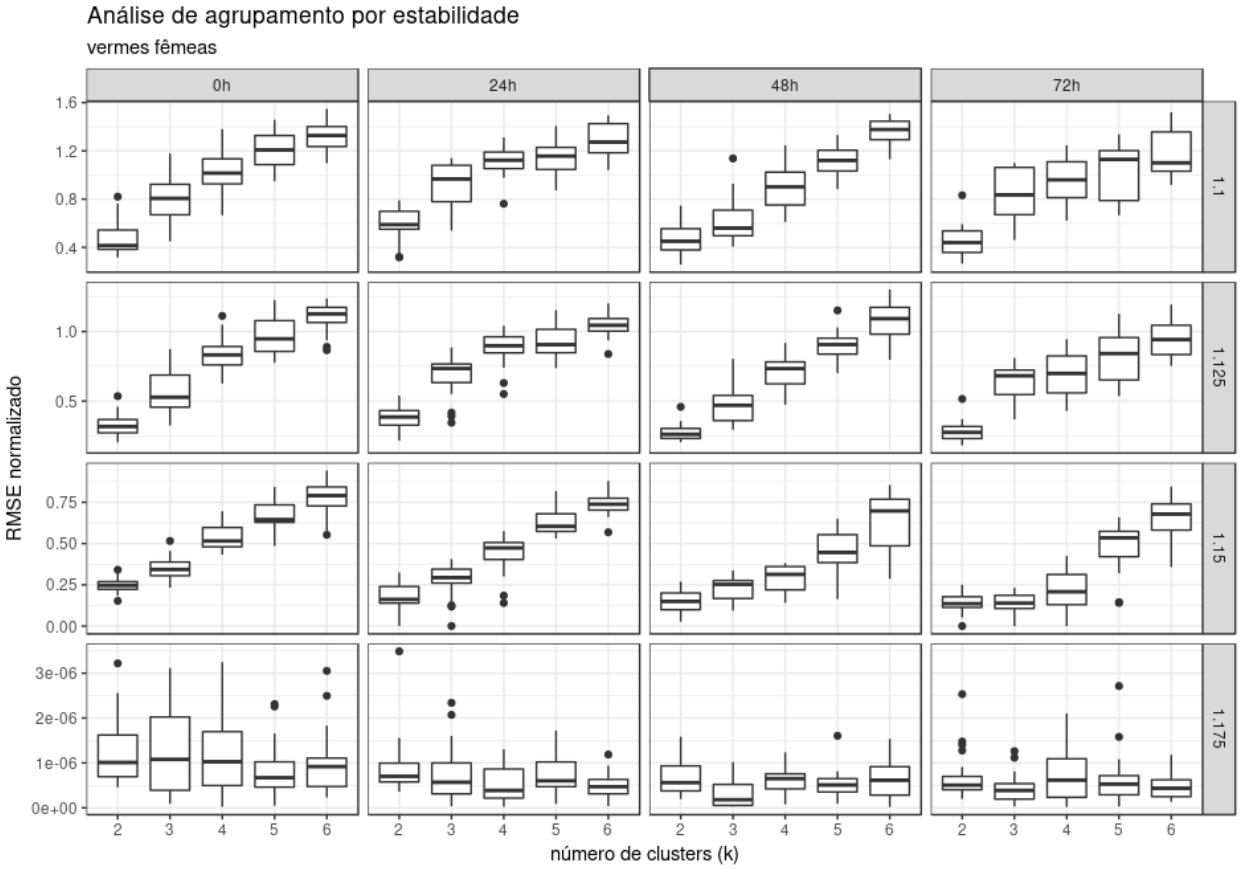


Figura 45: *Boxplots* para a análise de agrupamento por estabilidade para vermes fêmeas. Cada linha da tabela se refere a um valor de *membership exponent r.* O eixo das ordenadas (RMSE normalizado) se refere a razão entre o valor de RMSE calculado (para os conjuntos de dados de treinamento e de teste) e o RMSE calculado para coeficientes de pertencimento gerados pseudo-aleatoriamente.

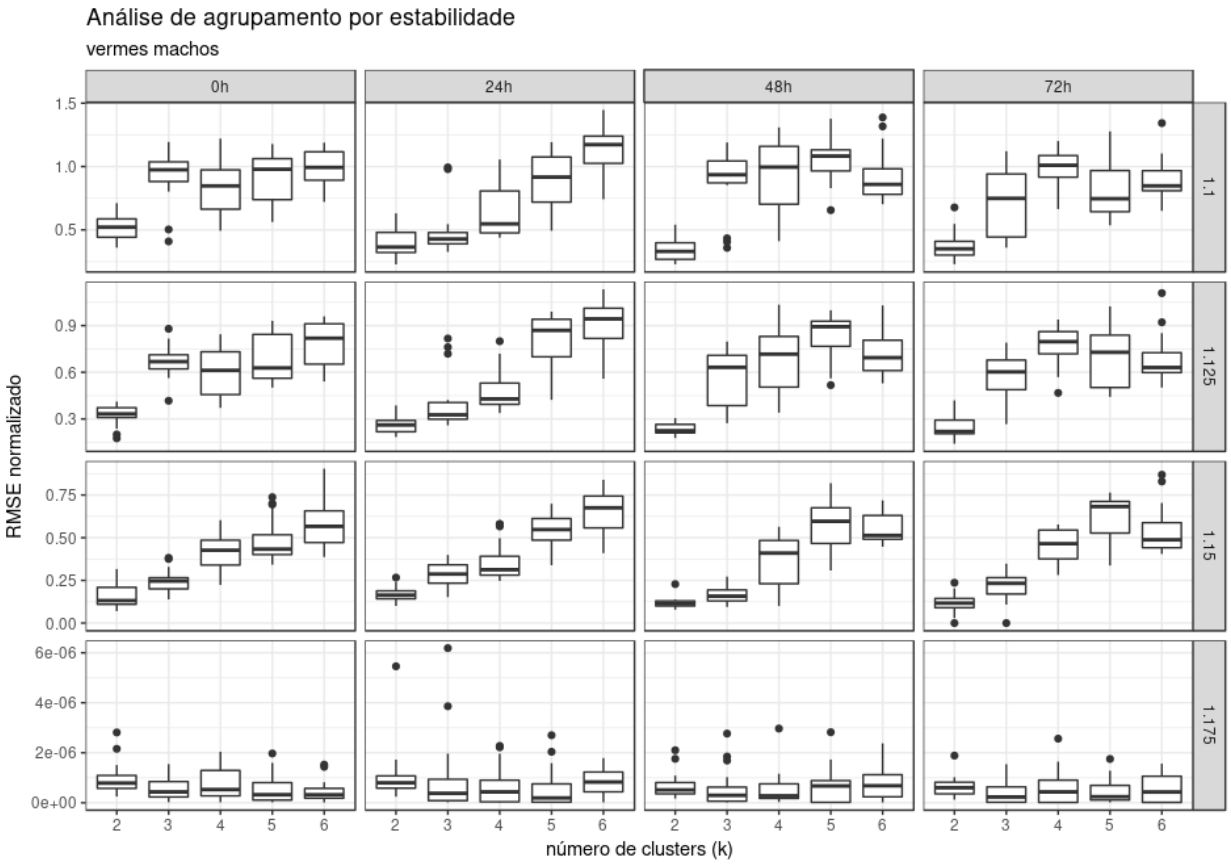


Figura 46: *Boxplots* para a análise de agrupamento por estabilidade para vermes machos. Cada linha da tabela se refere a um valor de *membership exponent r.* O eixo das ordenadas (RMSE normalizado) se refere a razão entre o valor de RMSE calculado (para os conjuntos de dados de treinamento e de teste) e o RMSE calculado para coeficientes de pertencimento gerados pseudo-aleatoriamente.

Os gráficos anteriores permitem avaliar qual a configuração de parâmetros que reduz o erro médio RMSE entre os resultados de agrupamento do conjunto de treinamento e do conjunto de teste. Para isso, não se deve comparar os valores de RMSE diretamente entre resultados com diferentes números de *clusters*, mas sim o RMSE normalizado, ou *corrigido por chance* (LANGE et al, 2004)[44]. Nota-se primeiramente que os resultados, em geral, foram independentes do sexo dos vermes e do tempo de tratamento. Além disso, nota-se que com o aumento do parâmetro *r* (que controla o grau de sobreposição entre os clusters) ocorre uma

[44] O RMSE corrigido (ou normalizado) deve ser definido - para estar em linha com a abordagem proposta por LANGE et al (2004) - como a razão entre o RMSE absoluto (calculado entre o conjunto de dados de treinamento e o cojunto de teste) e o RMSE entre o conjunto de teste e um conjunto de dados contendo coeficientes de pertencimento gerados pseudo-aleatoriamente. Isto é necessário porque, quanto maior o número de clusters adotado, menor será o RMSE absoluto calculado, mesmo para dados aleatórios, mostrando a necessidade de normalizar os valores de RMSE (ou outra medida de dscrepância que esteja sendo utilizada).

redução sistemática do RMSE normalizado. Os resultados obtidos para r = 1,175, entretanto, parecem corresponder a *clusters* tão difusos que já não se tem a capacidade de distinguir os resultados obtidos dos agrupamentos com diferentes números de *clusters*. Para os demais valores de *r* testados, é possível verificar uma variação da estabilidade em função do número de clusters *k*. A partir da observação das figuras 45 e 46, a configuração mais estável, para todos os tempos de tratamento, foi a partir dos resultados de agrupamento com expoente *r* = 1,15 ou 1,125 e *k* = 2 somente. Como a configuração *r* = 1,15 e k = 2 tem boa estabilidade (RMSE normalizado < 0,25), mas a diferença de estabilidade entre k=2 e k=3 foi pequena (∆RMSE normalizado ≈ 0,1), preferiu-se adotar a configuração com r = 1,125, que mostrou maior sensibilidade a mudanças no parâmetro k (∆RMSE normalizado ≈ 0,3).

O resultado k = 2 caracteriza a estrutura dos dados binária, indicando que o algoritmo de agrupamento utilizado teve a capacidade de distinguir entre 2 *clusters* de fenótipos somente. Eles foram considerados ser o *cluster* Controle (que contém a maioria dos vermes do grupo controle) e o *cluster não-Controle,* com vermes tratados e afetados pelo tratamento.

Uma expectativa deste trabalho era que a análise de agrupamento permitisse detectar não somente dois *clusters*, mas vários, detectando assim diferentes respostas fenotípicas. Acredita-se que este resultado *k* = 2 foi gerado pela inadequação do método de agrupamento testado, sugerindo que outros métodos deveriam ser testados a fim de avaliar a estrutura dos dados. Outra causa (improvável) é que conjunto de dados em si, utilizado neste trabalho, tenha pouca diversidade fenotípica. Supondo-se, entretanto, que a diversidade fenotípica não seja o problema, é possível que o conjunto de variáveis medidas nos quatro módulos (*CalculateImageOverlap, MeasureImageIntensity, MeasureTexture, MeasureGranularity*) não foram suficientes para detectar todas as caraterísticas fenotípicas relevantes. Dada a grande redundância de informação contida nas variáveis originais, como foi visto na análise exploratória e PCA, esta é uma hipótese plausível de ser feita.

Mesmo que a abordagem aqui empregada tenha levado a um provável resultado negativo na etapa da análise de agrupamento, como discutido acima, é interessante verificar qual o poder preditivo que o modelo árvore de regressão com efeitos aleatórios (*RE-EM tree*) pode obter utilizando-se os resultados de grau de

pertencimento ao *cluster* Controle. A tabela 5 resume alguns resultados iniciais obtidos. Comparando-se o erro quadrático médio RMSE do ajuste com a raiz quadrada da variância total estimada pelo modelo do tipo *RE-EM tree*, nota-se que o erro do ajuste é uma fração do erro global esperado pelo modelo, indicando que o ajuste foi razoável e que a abordagem testada neste trabalho pode ser um caminho para se obter um modelo preditor de compostos *hits,* mas precisa de mais estudos para julgar a sua competitividade com outras metodologias mais poderosas, porém mais complexas[45].

Tabela 5: Comparação entre o RMSE do ajuste pelo método *RE-EM tree* e os parâmetros estimados pelo modelo[46]

	Vermes Fêmeas				Vermes Machos			
tempo	0h	24h	48h	72h	0h	24h	48h	72h
σ_ε	0,518	0,350	0,256	0,286	0,361	0,318	0,287	0,285
σ_{RE}	0,496	0,317	0,554	0,478	0,317	0,478	0,490	0,453
$\sqrt{(\sigma_\varepsilon^2 + \sigma_{RE}^2)}$	0,717	0,472	0,611	0,557	0,480	0,574	0,568	0,535
RMSE	0,412	0,384	0,108	0,147	0,272	0,174	0,150	0,151

[45] E que, consequentemente, geram resultados mais difíceis de serem interpretados.

[46] O parâmetro σ_ε^2 é variância estimada do erro e o parâmetro σ_{RE}^2 é a variância estimada dos efeitos aleatórios a nível de verme presentes em um modelo REEMtree. A soma $\sigma_\varepsilon^2 + \sigma_{RE}^2$ é uma estimativa da variância global do modelo do tipo *RE-EM tree*. Os parâmetros σ_e e σ_{RE} são consequentemente o desvio padrão do erro e dos efeitos aleatórios.

5 CONCLUSÕES E PERSPECTIVAS

A partir da análise estatística exploratória dos dados pôde-se propor e comparar modelos para a inferência do efeito de cada composto testado, a partir de dados de motilidade do parasita. Esta análise permitiu obter informações, tal que modelos lineares **generalizados multinível** foram considerados necessários para se ter adequação à estrutura hierárquica e longitudinal dos dados. O melhor modelo univariado proposto neste trabalho foi aquele que assume uma distribuição Poisson Generalizada, onde efeitos aleatórios a nível de verme foram necessários para se obter um bom ajuste aos dados. O modelo SEM para modelar estes efeitos de heterogeneidade foi particularmente eficaz.

Para tornar o modelo mais robusto, foram comparadas duas propostas de inclusão da informação de similaridade química entre os compostos, utilizando o conceito de espaço químico representado por uma rede. Com os dados testados, não foi possível notar grandes vantagens no uso destas abordagens quando comparado ao modelo mais simples, que não considera a estrutura química dos compostos explicitamente. Pôde mostrar, entretanto, que o aumento da complexidade não foi danoso, como mostra a comparação dos resultados de WAIC. É esperado, entretanto, que vantagens possam ser notadas em futuras aplicações do modelo a dados de experimentos com compostos de uma série congênere.

A adaptação dos modelos multinível para aplicação no estudo dos resultados de experimentos do tipo Dose-Resposta foi bem-sucedida, mas não foi possível neste trabalho produzir modelos que permitam a estimação dos parâmetros de interesse de uma forma unificada entre as várias curvas em cada tempo estudado. Isto provavelmente melhoraria a estimativa destes parâmetros.

A análise multivariada foi desenvolvida em busca de um modelo de classificação de compostos considerados *hits*, a partir da análise fenotípica mais ampla dos parasitas. Para isso, a abordagem proposta e testada foi: a transformação, redução de dimensionalidade e agrupamento *fuzzy* dos dados, concluindo em um modelo preditivo do tipo árvore de regressão com efeitos aleatórios. A etapa de agrupamento dos dados não produziu resultados que caracterizassem o espaço fenotípico em um bom nível de detalhe. Isto limitou a

análise restante, mas acredita-se que o modelo multivariado resultante é um protótipo promissor, que carece de mais desenvolvimento.

É de interesse avançar no desenvolvimento dos modelos univariados propostos, avaliando a possiblidade de melhor aproveitamento de rede de compostos, *scaffolds* e *sub-scaffolds* na modelagem de experimentos de triagem. Como alternativas temos a aplicação de medidas de rede, ou construir uma versão generalizada destas redes. Espera-se poder aplicar os *scripts* gerados em novos conjuntos de dados para avaliar a vantagem dos modelos de rede em comparação ao modelo básico.

Faz-se necessário verificar as hipóteses levantadas na discussão dos resultados sobre como melhorar o planejamento experimental; persistindo o problema, buscar modificar o modelo existente para melhorá-lo, incluindo ajuste para dados censurados à direita.

Uma última perspectiva é a implementação destes modelos univariados na forma de um *R-package* para disponibilizar os *scripts* escritos incluindo etapas de tratamento dos dados, geração das redes, ajuste dos modelos e ferramentas para a análise dos resultados.

Na análise multivariada, é importante estudar outros algoritmos de agrupamento *fuzzy* que possam detectar uma maior diversidade de fenótipos, inclusive com novos dados experimentais que contenham uma grande diversidade de tratamentos.

6 REFERÊNCIAS BIBLIOGRÁFICAS

AITCHISON, J. *The statistical analysis of compositional data*. ***Journal of the Royal Statistical Society***, Edinburgh, v. 44, 139-177, 1982.

ALBRECHT, D.R.; BARGMANN, C.I. *High-content behavioral analysis of Caenorhabditis elegans in precise spatiotemporal chemical environments*. ***Nature Methods***, v. 8, p. 599-605, 2011.

ARAGON, A.D. *et al*. *Towards an understanding of the mechanism of action of praziquantel.* ***Molecular and Biochemical Parasitology***, v. 164, p.57–65, 2009.

ASARNOW, D. *et al*. *The QDREC web server: determining dose–response characteristics of complex macroparasites in phenotypic drug screens.* ***Bioinformatics****,* v. 31(9), p. 1515–1518, 2015.

ASARNOW, D.; SINGH, R. *Segmenting the etiological agent of schistosomiasis for high-content screening*. ***IEEE Transactions on Medical Imaging***, Febuary 2013, pp. 1007-1018.

BAI, X. *et al*. *Splitting touching cells based on concave points and ellipse fitting*. ***Pattern Recognition****,* v. 42, p.2434-2446, 2009.

BARABÁSI, A. L.; *Linked: how everything is connected to everything else and what it means for business, science, and everyday life*. PLUME, Penguin Books, New York, 2003.

BATES, D. *et al. Fitting Linear Mixed-Effects Models Using lme4.* ***Journal of Statistical Software****, 67(1), 1-48, 2015.*

BESAG, J.; KOOPERBERG, C. ***Biometrika***, 82, 733, 1995.

BESAG, J.; YORK, J.; MOLLIÉ, *A. Bayesian image restoration, with two applications in spatial statistics.* ***Annals of the Institute of Statistical Mathematics****.* 43(1): 1–20, 1991.

BIVAND, R. S.; GÓMEZ-RUBIO, V.; RUE, H. *Approximate Bayesian inference for spatial econometrics models.* ***Spatial Statistics***, v. 9, p. 146-165, 2014.

BREIMAN, L. *Random forests*. ***Machines Learning***. 45, p. 5-32, 2001.

BREIMAN, L. *et al*. CART: *Classification and Regression Trees. Belmont*, CA: Wadsworth, 1983.

BUCKINGHAM, S. D.; PARTRIDGE, F. A.; SATTELLE. D. B. Automated, high-throughput, motility analysis in Caenorhabditis elegans and parasitic nematodes: Applications in the search for new anthelmintics. ***Int. J. Parasitol. Drugs Drug. Resist.***, 4, 226, 2014.

BUCHSER, W. et al. *Assay development guidelines for image-based high content screening, high content analysis and high content imaging. In Assay Guidance Manual* Bethesda, **National Center for Advancing Translational Sciences**. 2004.

BULLEN, A. *Microscopic imaging techniques for drug discovery*. ***Nature Reviews Drug Discovery***, v. 7, p. 54-67, 2008.

BUSSAB, W. O. e MORETTIN, P. A. Estatística Básica. 9ª Edição, Editora Saraiva, 2017.

CAFFREY, C. R. *Schistosomiasis and its treatment.* ***Future Medicinal Chemistry***, v. 7, n. 6, p. 675–676, 2015.

CARHART, R. E.; SMITH, D. H.; VENKATARAGHAVAN, R. *Atom Pairs as Molecular Features in Structure-Activity Studies: Definition and Applications*. ***Journal of Chemical Information and Computer Sciences***, 25, 64-73, 1985.

CARPENTER, A.E. *et al*. *CellProfiler: image analysis software for identifying and quantifying cell phenotypes*. ***Genome Biology***, v. 7, R100, 2006.

CARPENTER, A.E. e JONES, T. R. *CellProfiler cell image analysis software.* versão 2.0. Disponível em: <https://cellprofiler.org/manuals/>. Acesso em: 20 de dezembro de 2018.

CASTAÑÓN, C.A.B. Análise e reconhecimento digital de formas biológicas para o diagnóstico automático de parasitas do gênero Eimeria. 2006. 161 f. Tese (Doutorado em Bioinformática) - Universidade de São Paulo, São Paulo, 2006.

CEDERGREEN, N.; RITZ, C.; STREIBIG, J. C. *Improved Empirical Models Describing Hormesis.* ***Environmental Toxicology and Chemistry****,* v.24, 12, p. 3166-3172, 2005.

CHEN, X. e REYNOLDS, C. H. *Performance of Similarity Measures in 2D Fragment-Based Similarity Searching: Comparison of Structural Descriptors and Similarity Coefficients.* ***Journal of Chemical Information and Computer Sciences****, 42 (6), pp 1407–1414, 2002.*

CIOLI, D. *et al*. *Schistosomiasis control: praziquantel forever?* ***Molecular and Biochemical Parasitology***, v. 195, n. 1, p. 23–29, 2014.

COLLEY, D. G. *et al*. *Human schistosomiasis.* ***The Lancet***, v. 383, n. 9936, p. 2253–2264, jun. 2014.

CONERY, A.L. *et al*. *High-throughput screening for novel anti-infectives using a C. elegans pathogenesis model*. ***Current Protocols in Chemical Biology***, v. 6, p. 25-37, 2014.

CONSUL, P. C. e JAIN, G. C. *A Generalization of the Poisson Distribution*, ***Technometrics,*** v. 15:4, 791-799, 1973.

CRONIN, C.J. *et al. An automated system for measuring parameters of nematode sinusoidal movement*. ***BMC Genetics****, v.6,* 2005.

CRUZ, D. J. M.; KOISHI, A. C., TANIGUCHI, J. B., LI, X., MILAN BONOTTO, R., No, J. H., FREITAS-JUNIOR, L. H. *High content screening of a kinase-focused library reveals compounds broadly-active against dengue viruses*. ***PLOS Neglected Tropical Diseases***. v. 7, n. 2, e. 2073, 2013.

DALRYMPLE, M. L.; HUDSON, I. L.; FORD, R. P. K. *Finite Mixture, Zero-inflated Poisson and Hurdle models with application to SIDS.* ***Computational Statistics & Data Analysis***, v. 41, p. 491-504, 2003.

DOBSON, C. M. *Chemical space and biology*. ***Nature***, 432:824–828, 2004.

EL-LAKKANY, N.M. e SEIF EL-DIN, S.H. *Haemin enhances the in vivo efficacy of artemether against juvenile and adult Schistosoma mansoni in mice.* ***Parasitology Research****,* 112: 2005, 2013.

FREES, E. W. *Regression Modeling with Actuarial and Financial Applications*. ***Cambridge University Press***, 2011.

FISKER, R., *Making deformable template models operational*. ***Technical University of Denmark***, *Lyngby*, 2000.

GABOR, D. *Theory of communication.* ***Journal of the Institute of Electrical Engineers****,* 93:429-441, 1946.

GAMERMAN, D; LOPES, H. F. Markov Chain Monte Carlo: *Stochastic Simulation for Bayesian Inference*. ***Chapman & Hall/CRC - Taylor & Francis Group***, 2ª Ed., 2006.

GELMAN, A.; HILL, J. *Data Analysis Using Regression and Multilevel/Hierarchical Models*. ***Cambridge University Press***, 1ª Ed., 2007.

GELMAN, A; HWANG, J.; VEHTARI, A. *Understanding predictive information criteria for Bayesian models*. ***Statistics and Computing***, v. 24, 6, 997–1016, 2014.

GENG, W. *et al*. *Automatic tracking, feature extraction and classification of C. elegans phenotypes*. ***IEEE Transactions on Biomedical Engineering***, v. 51, p.1811-1820, 2004.

GRAVES, R. *Incorporating Transmitted Light Modalities into High-Content Analysis Assays.* ***In Cooper M, Mayr LM (eds): Label-Free Technologies for Drug Discovery****. Chichester, John Wiley & Sons*, Ltd, 2011.

HARALICK, R. M.; SHANMUGAM, K.; DINSTEIN, I. *Textural Features for Image Classification.* ***IEEE Transaction on Systems Man, Cybernetics***, SMC-3(6):610-621, 1973.

HARTIG, F. (2018). *DHARMa: Residual Diagnostics for Hierarchical (Multi-Level / Mixed) Regression Models*. R package version 0.1.6. https://CRAN.R-project.org/package=DHARMa

HOTEZ, P. J. *et al*. *Helminth Infections: Soil-transmitted Helminth Infections and Schistosomiasis.* ***Disease Control Priorities in Developing Countries***, p. 467–482, 2006.

HORAN, K. et al (2018). *ChemmineR: Cheminformatics Toolkit for R.* R package version 2.28.3. https://github.com/girke-lab/ChemmineR

HUANG, K. *et al. Using articulated models for tracking multiple C. elegans in physical contact*. ***Journal of Signal Processing Systems***, v. 55, p.113-126, 2009.

HUBERT, L.; ARABIE, P. *Comparing partitions.* ***Journal of Classification***, 193–218, 1985.

INGRAM, K.; ELLIS, W.; KEISER, J*. Antischistosomal Activities of Mefloquine-Related Arylmethanols.* ***Antimicrobial Agents and Chemotherapy***, 56 (6) 3207-3215, 2012.

JAMES, G. et al; *An Introduction to Statistical Learning: with Applications in R.* ***Springer Texts in Statistics***. 2013.

JOE, H. e ZHU, R. *Generalized Poisson Distribution: the Property of Mixture of Poisson and Comparison with Negative Binomial Distribution.* ***Biometrical Jounal***, 47(2), p. 219-229, 2005.

JONES T. R. *et al*. *CellProfiler Analyst: data exploration and analysis software for complex image-based screens*. ***BMC Bioinformatics***, v. 9, p. 482, 2008.

JONES T. R. *et al*. *Scoring Diverse Cellular Morphologies in Image-Based Screens with Iterative Feedback and Machine Learning*. ***Proceedings of the National Academy of Sciences***, v. 106, p.1826-1831, 2009.

KAUFMAN, L. e ROUSSEEUW, P. J. *Finding Groups in Data: An Introduction to Cluster Analysis*. Volume 344 de *Wiley Series in Probability and Statistics*. Ed. 99, ***John Wiley & Sons****, 2009.*

KIMMEL, J. C.; BRACK, A. S.; MARSHALL, W. F. *Deep convolutional and recurrent neural networks for cell motility discrimination and prediction.* ***Pre-print*** **em** ***bioRxiv****,* disponível *online* desde 3 de Julho de 2017.
doi: http://dx.doi.org/10.1101/159202.

LAMBERT, D. *Zero-inflated Poisson regression with an application to defects in manufacturing.* ***Technometrics***, v. 34(1), p. 1-14, 1992.

LAMPRECHT, M.R. *et al. CellProfiler: free, versatile software for automated biological image analysis****. Biotechniques***, v. 42, p. 71–75, 2007.

LANGE T. et al. *Stability-based validation of clustering solutions.* ***Neural Comput.*** 16(6):1299–323, 2004.

LEISCH, F. *A Toolbox for K-Centroids Cluster Analysis.* ***Computational Statistics and Data Analysis***, v. 51 (2), 526-544, 2006.

LEE, H. *et al. Quantification and clustering of phenotypic screening data using time-series analysis for chemotherapy of schistosomiasis*. ***BMC Genomics***, v.13, Suppl. 1, 2012.

LI, H. J. et al. *Dihydroartemisinin-praziquantel combinations and multiple doses of dihydroartemisinin in the treatment of Schistosoma japonicum in experimentally infected mice.* ***Annals of tropical medicine and parasitology.*** v. 105, 4: p.329-33, 2011.

LIN, G. *et al. A multi-model approach to simultaneous segmentation and classification of heterogeneous populations of cell nuclei in 3D confocal microscope images*. ***Cytometry A***, v. 71, p. 724-736, 2007.

LINDGREN, F.; RUE, H.; LINDSTROM, J. *An Explicit Link Between Gaussian Fields and Gaussian Markov Random Fields: The Stochastic Partial Differential Equation Approach (with discussion).* ***Journal of the Royal Statistical Society B***, 73(4), 423-498, 2011.

LINDGREN, F.; RUE, H. *Bayesian Spatial Modelling with R-INLA*. ***Journal of Statistical Software***, 63(19), 1-25, 2015. URL http://www.jstatsoft.org/v63/i19/.

LJOSA, V.; CARPENTER, A.E. *Introduction to the quantitative analysis of two-dimensional fluorescence microscopy images for cell-based screening*. ***PLoS Computational Biology***, v. 5, 2009.

LLOYD-SMITH, J. O*. Maximum Likelihood Estimation of the Negative Binomial Dispersion Parameter for Highly Overdispersed Data, with Applications to Infectious Diseases. **PLoS ONE**, 2 (2): e180,* 2007.

MARCELLINO, C. *et al. WormAssay: a novel computer application for whole-plate motion-based screening of macroscopic parasites*. ***PLoS Neglected Tropical Diseases***, v.6, 2012.

MAGGIORA, G. M.; BAJORATH, J. J. *Chemical space networks: a powerful new paradigm for the description of chemical space*. ***Comput. Aided. Mol. Des.***, 28, 795, 2014.

MATHAI A. M.; SAXENA, R. K. *On a generalized hypergeometric distribution.* ***Metrika***, v. 11(1), 127-132, 1967.

MAECHLER, M. et al (2018). *Cluster: "Finding Groups in Data": Cluster Analysis Extended Rousseeuw et al*. R package version 2.0.7-1. https://CRAN.R-project.org/package=cluster

MAHALANOBIS, P. C. *On the Generalized Distance in Statistics.* ***Proceedings of the National Institute of Science of India.*** 2, 49–55

MATLOCK, M. K.; ZARETZKI, J. M.; SWAMIDASS, S. J. *Scaffold network generator: a tool for mining molecular structures*. ***Bioinformatics***, 29, 2655, 2013.

MARTINS, T. G.; SIMPSON, D; LINDGREN, F; RUE, H. *Bayesian computing with INLA: New features.* ***Computational Statistics and Data Analysis***, 67, 68-83, 2013.

MEGASON, S.G.; FRASER, S.E. *Imaging in systems biology*. ***Cell***, v.130, p. 784-795, 2007.

MELO-FILHO, C. C *et al. QSAR-Driven Discovery of Novel Chemical Scaffolds Active Against Schistosoma Mansoni.* ***Journal of Chemical Information and Modeling***, v. 56 (7), p. 1357–1372, 2016.

MENDONÇA-SILVA, D. L. et al. *Characterization of a GABAergic neurotransmission in adult Schistosoma mansoni.* ***Parasitology***, v. 129(2), p.1-10, 2004.

MONÉ, H.; BOISSIER, J. *Sexual Biology of Schistosomes. In:* ***Advances in Parasitology***, v. 57, p. 89–189.

MITSUI, Y.; MIURA, M.; AOKI, Y. *In vitro effects of artesunate on the survival of worm pairs and egg production of Schistosoma mansoni.* ***Journal of Helminthology***, 83(1), 7-11, 2009.

MULLAHY, J. *Specification and testing of some modified count data models.* ***Journal of Econometrics***, v. 33, p. 341-365, 1986;

NARE, B.; SMITH, J.M.; PRICHARD, R.K. *Differential effects of oltipraz and its oxy-analogue on the viability of Schistosoma mansoni and the activity of glutathione-s-transferase****. Biochem. Pharmacol,*** v.42, p.1287-1292, 1991.

NEVES, B. J. *et al*. *Discovery of New Anti-Schistosomal Hits by Integration of QSAR-Based Virtual Screening and High Content Screening.* ***Journal of Medicinal Chemistry***, v. 59(15), p. 7075-88, 2016a.

NEVES, B. J. *et al. The antidepressant drug paroxetine as a new lead candidate in schistosome drug discovery.* ***Medicinal Chemistry Communication - Royal Society of Chemistry***, v. 7, p. 1176-1182, 2016b.

NOEL, F.; MENDONÇA-SILVA, D.L.; THIBAUT, J.P.; LOPES, D.V. *Characterization of two classers of benzodiazepine binding sites in Schistosoma mansoni.* ***Parasitology****,* v.22: 1-10, 2007.

NORTON, E. C.; DOWD, B. E. *Log Odds and the Interpretation of Logit Models.* ***Health Services Research****, 53(1), 2017.*

PANIC, G. *Activity Profile of an FDA-Approved Compound Library against Schistosoma mansoni.* ***PLOS Neglected Tropical Diseases****,* v.9(7): e0003962, 2015.

PAVELEY, R. A.; BICKLE, Q. D. *Automated Imaging and other developments in whole-organism anthelmintic screening****. Parasite Immunology***, v. 35, p. 302–313*,* 2013.

PAVELEY, R. A. *et al*. *Whole organism high-content screening by label-free, image-based Bayesian classification for parasitic diseases.* ***PLOS Neglected Tropical Diseases****,* v. 6, n. 7, e 1762, 2012.

PEAK, E.; CHALMERS, I. W.; HOFFMANN, K. F. *Development and Validation of a Quantitative, High-Throughput, Fluorescent-Based Bioassay to Detect Schistosoma Viability.* ***PLOS Neglected Tropical Diseases****,* 4(7): e759, 2010.

PERLMAN Z. E. *et al. Multidimensional Drug profiling by Automated Microscopy*. ***Science***, v. 306, p. 1194-1198, 2004.

R Core Team (2016). *R: A language and environment for statistical computing. R Foundation for Statistical Computing*, Vienna, Áustria.
URL http://www.R-project.org/.

RAMIREZ, B. *et al*. *Schistosomes: challenges in compound screening.* ***Expert Opinion on Drug Discovery***, v. 2, p. S53–S61, 2007.

RAMIREZ-LOPEZ, L. e STEVENS, A. (2016). *resemble: Regression and Similarity Evaluation for Memory-Based Learning in Spectral Chemometrics.* R package version 1.2.2. https://CRAN.R-project.org/package=resemble

RAND, W. M. *Objective criteria for the evaluation of clustering methods.* ***Journal of the American Statistical Association***, v. 66 (336), 846–850, 1971.

RITZ, C.; BATY, F.; STREIBIG, J.C.; GERHARD, D. *Dose-Response Analysis Using R. PLOS ONE,* 10(12): e 0146021, 2015.

ROSS, A. G. P. *et al*. *Schistosomiasis.* ***New England Journal of Medicine***, v. 346, n. 16, p. 1212–1220, 18 abr. 2002.

ROUSSEL, N. *et al. A computational model for C. Elegans locomotory behavior: application to multiworm tracking*. ***IEEE Transactions in Biomedical Engineering***, v. 54, p. 1786-1797, 2007.

RUE, H.; MARTINO, S.; CHOPIN, N. *Approximate Bayesian Inference for Latent Gaussian Models Using Integrated Nested Laplace Approximations (with discussion)*. ***Journal of the Royal Statistical Society***, Series B, v. 71, p. 319–392, 2009.

RUE, H *et al*. *Bayesian computing with INLA: A review.* ***Annual Reviews of Statistics and Its Applications***, v.4, 395-421, 2017. URL http://arxiv.org/abs/1604.00860

SELA, R. J. e SIMONOFF, J. S. *RE-EM trees: a data mining approach for longitudinal and clustered data.* ***Machine Learning****,* v. 86, p. 169-207, 2012.

SELA, R. J. e SIMONOFF, J. S. (2015). *REEMtree: Regression Trees with Random Effects for Longitudinal (Panel) Data.* R package version 0.90.3. https://CRAN.R-project.org/package=REEMtree

SIMPSON, D. P. *et al. Penalising model component complexity: A principled, practical approach to constructing priors. arXiv preprint arXiv:1403.4630, 2014.*

SINGH, R. *et al. Automated Image-Based Phenotypic Screening for High-Throughput Drug Discovery*. ***IEEE International Symposium on Computer-Based Medical Systems****,* 2009.

SINGH, R.; BEASLEY, R.; LONG, T.; CAFFREY, C.R. *Algorithmic Mapping and Characterization of the Drug-Induced Phenotypic-Response Space of Parasites Causing Schistosomiasis*. ***IEEE/ACM transactions on computational biology and bioinformatics***, 2016.

SINGH, R. *Quantitative High-Content Screening-Based Drug Discovery against Helmintic Diseases*, *in Parasitic Helminths: Targets, Screens, Drugs, and Vaccines*, **Ed. C. Caffrey, Wiley-Blackwell**, 2012, p.159-179.

SOMMER, C. e GERLICH, D.W. *Machine learning in cell biology – teaching computers to recognize phenotypes.* ***Journal of Cell Science****, v. 126, p. 5529–5539, 2013.*

STARKUVIENE, V.; PEPPERKOK, R. *The potential of high-content high-throughput microscopy in drug Discovery*. ***British Journal of Pharmacology,*** v.152, p. 62-71, 2007.

STEPHENS, G.J. *et al. Dimensionality and dynamics in the behavior of C. elegans*. ***PLoS Computational Biology***, 2008.

TANAKA, M. *et al. An Unbiased Cell Morphology-Based Screen for New Biologically Active Small Molecules*. ***PLoS Biology***, 2005.

TIAN, Y. *et al. C. elegans screen identifies autophagy genes specific to multicellular organisms*, ***Cell***, v.141, p. 1042-1055, 2010.

UHLENBECK, G. E.; ORNSTEIN, L. S. *On the theory of Brownian Motion*. ***Physical Review Journals Archive,*** 36: 823–841, 1930.

VARADHAN, R. (2015). *alabama: Constrained Nonlinear Optimization*. R package version 2015.3-1. *https://CRAN.R-project.org/package=alabama*

VARIN, T. et al. *Mining for bioactive scaffolds with scaffold networks: improved compound set enrichment from primary screening data*. ***Jounal of Chemical Information and Modeling***. 51(7):1528-38, 2011.

WAHLBY, C. *et al. An image analysis toolbox for high-throughput C. elegans assays.* ***Nature Methods***, v. 9, p. 714-716, 2012.

WAHLBY, C. *et al. Resolving clustered worms via probabilistic shape models*. In ***IEEE International Symposium on Biomedical Imaging: From Nano to Macro***, Rotterdam, The Netherlands, June 2010, p. 552–555.

WANG, W; WANG, L; LIANG, Y. S. *Susceptibility or resistance of praziquantel in human schistosomiasis: a review.* ***Parasitology Research***, v. 111(5), p. 1871–1877, 2012.

ZAMANI, H. e ISMAIL, N. *Functional Form for the Generalized Poisson Regression Model.* ***Communications in Statistics - Theory and Methods***, 41:20, 3666-3675, 2012.

ZANELLA, F. *et al*. *High content screening: seeing is believing.* ***Trends in Biotechnology***, v. 28, p. 237–45, 2010.

7 ANEXOS

7.1 Anexo 1: *Script* para o *pipeline* de processamento de imagens em *software* CellProfiler

A letra reduzida é para economia de espaço.

CellProfiler Pipeline: http://www.cellprofiler.org
Version:3
DateRevision:20160503183100
GitHash:ac0529e
ModuleCount:41
HasImagePlaneDetails:False

Images:[module_num:1|svn_version:\'Unknown\'|variable_revision_number:2|show_window:True|notes:\x5B\'To begin creating your project, use the Images module to compile a list of files and/or folders that you want to analyze. You can also specify a set of rules to include only the desired files in your selected folders.\', \'---\', \'Drag and drop the ExampleIlluminationCorrection folder into the File list panel.\'\x5D|batch_state:array(\x5B\x5D, dtype=uint8)|enabled:True|wants_pause:False]
    :
    Filter images?:Custom
    Select the rule criteria:and (extension does istif) (file doesnot contain "Thumb")

Metadata:[module_num:2|svn_version:\'Unknown\'|variable_revision_number:4|show_window:True|notes:\x5B\'The Metadata module optionally allows you to extract information describing your images (i.e, metadata) which will be stored along with your measurements. This information can be contained in the file name and/or location, or in an external file.\'\x5D|batch_state:array(\x5B\x5D, dtype=uint8)|enabled:True|wants_pause:False]
    Extract metadata?:Yes
    Metadata data type:Text
    Metadata types:{}
    Extraction method count:4
    Metadata extraction method:Extract from file/folder names
    Metadata source:File name
    Regular expression:^.*?(?P<Well>\x5BA-P\x5D\\\\d{2}).TIF$
    Regular expression:(?P<Date>\x5B0-9\x5D{4}_\x5B0-9\x5D{2}_\x5B0-9\x5D{2})$
    Extract metadata from:All images
    Select the filtering criteria:and (file does contain "")
    Metadata file location:
    Match file and image metadata:\x5B\x5D
    Use case insensitive matching?:No
    Metadata extraction method:Extract from file/folder names
    Metadata source:Folder name
    Regular expression:^(?P<Plate>.*)_(?P<Well>\x5BA-P\x5D\x5B0-9\x5D{2})_s(?P<Site>\x5B0-9\x5D)_w(?P<ChannelNumber>\x5B0-9\x5D)
    Regular expression:^.*?\x5BPp\x5Dlate_(?P<Plate>.*)/TimePoint_(?P<Number>\\\\d+)$
    Extract metadata from:All images
    Select the filtering criteria:and (file does contain "")
    Metadata file location:
    Match file and image metadata:\x5B\x5D
    Use case insensitive matching?:No

Metadata extraction method:Import from file

Metadata source:File name

Regular expression:^(?P<Plate>.*)_(?P<Well>\x5BA-P\x5D\x5B0-9\x5D{2})_s(?P<Site>\x5B0-9\x5D)_w(?P<ChannelNumber>\x5B0-9\x5D)

Regular expression:(?P<Date>\x5B0-9\x5D{4}_\x5B0-9\x5D{2}_\x5B0-9\x5D{2})$

Extract metadata from:All images

Select the filtering criteria:and (file does contain "")

Metadata file location:/data/bioensaios/Users/floriano/Schisto_drug_tests-07-03-16/firstlast.csv

Match file and image metadata:\x5B{\'Image Metadata\'\x3A u\'Plate\', \'CSV Metadata\'\x3A u\'Plate\'}, {\'Image Metadata\'\x3A u\'Well\', \'CSV Metadata\'\x3A u\'Well\'}, {\'Image Metadata\'\x3A u\'Number\', \'CSV Metadata\'\x3A u\'Number\'}\x5D

Use case insensitive matching?:No

Metadata extraction method:Import from file

Metadata source:File name

Regular expression:^(?P<Plate>.*)_(?P<Well>\x5BA-P\x5D\x5B0-9\x5D{2})_s(?P<Site>\x5B0-9\x5D)_w(?P<ChannelNumber>\x5B0-9\x5D)

Regular expression:(?P<Date>\x5B0-9\x5D{4}_\x5B0-9\x5D{2}_\x5B0-9\x5D{2})$

Extract metadata from:All images

Select the filtering criteria:and (file does contain "")

Metadata file location:/data/bioensaios/Users/floriano/Schisto_drug_tests-07-03-16/Metadados_07_03_16.csv

Match file and image metadata:\x5B{\'Image Metadata\'\x3A u\'Plate\', \'CSV Metadata\'\x3A u\'Plate\'}, {\'Image Metadata\'\x3A u\'Well\', \'CSV Metadata\'\x3A u\'Well\'}\x5D

Use case insensitive matching?:No

NamesAndTypes:[module_num:3|svn_version:\'Unknown\'|variable_revision_number:6|show_window:True|notes:\x5B\'The NamesAndTypes module allows you to assign a meaningful name to each image by which other modules will refer to it.\', \'---\', \'The rule criteria will select only one file from the full list\x3A ADSAStaphInfection2_A01_w2247376DD-6ADD-442D-AE47-F54A05F3EA94.tif\'\x5D|batch_state:array(\x5B\x5D, dtype=uint8)|enabled:True|wants_pause:False]

Assign a name to:Images matching rules

Select the image type:Grayscale image

Name to assign these images:OrigWorms

Match metadata:\x5B{u\'NextOrigWorms\'\x3A u\'Plate\', u\'PrevOrigWorms\'\x3A u\'Plate\'}, {u\'NextOrigWorms\'\x3A u\'Well\', u\'PrevOrigWorms\'\x3A u\'Well\'}, {u\'NextOrigWorms\'\x3A u\'PrevNumber\', u\'PrevOrigWorms\'\x3A u\'NextNumber\'}\x5D

Image set matching method:Metadata

Set intensity range from:Image metadata

Assignments count:2

Single images count:0

Maximum intensity:255.0

Select the rule criteria:and (metadata does Last "0")

Name to assign these images:PrevOrigWorms

Name to assign these objects:Cell

Select the image type:Grayscale image

Set intensity range from:Image metadata

Retain outlines of loaded objects?:No

Name the outline image:LoadedObjects

Maximum intensity:255.0

Select the rule criteria:and (metadata does First "0")

Name to assign these images:NextOrigWorms

Name to assign these objects:Nucleus

Select the image type:Grayscale image

Set intensity range from:Image metadata

Retain outlines of loaded objects?:No

Name the outline image:LoadedOutlines

Maximum intensity:255.0

Groups:[module_num:4|svn_version:\'Unknown\'|variable_revision_number:2|show_window:True|notes:\x5B\'The Groups module optionally allows you to split your list of images into image subsets (groups) which will be processed independently of each other. Examples of groupings include screening batches, microtiter plates, time-lapse movies, etc.\'\x5D|batch_state:array(\x5B\x5D, dtype=uint8)|enabled:True|wants_pause:False]

Do you want to group your images?:Yes

grouping metadata count:2

Metadata category:Plate

Metadata category:Well

ImageMath:[module_num:5|svn_version:\'Unknown\'|variable_revision_number:4|show_window:False|notes:\x5B\'Invert the intensity of the image, since the background method in CorrectIlluminationCalc assumes a light foreground and dark background.\'\x5D|batch_state:array(\x5B\x5D, dtype=uint8)|enabled:True|wants_pause:False]

Operation:Invert

Raise the power of the result by:1

Multiply the result by:1

Add to result:0

Set values less than 0 equal to 0?:Yes

Set values greater than 1 equal to 1?:Yes

Ignore the image masks?:No

Name the output image:PrevInvertedWorms

Image or measurement?:Image

Select the first image:PrevOrigWorms

Multiply the first image by:1

Measurement:

Image or measurement?:Image

Select the second image:

Multiply the second image by:1

Measurement:

ImageMath:[module_num:6|svn_version:\'Unknown\'|variable_revision_number:4|show_window:False|notes:\x5B\x5D|batch_state:array(\x5B\x5D, dtype=uint8)|enabled:True|wants_pause:False]

Operation:Invert

Raise the power of the result by:1

Multiply the result by:1

Add to result:0

Set values less than 0 equal to 0?:Yes

Set values greater than 1 equal to 1?:Yes

Ignore the image masks?:No

Name the output image:NextInvertedWorms

Image or measurement?:Image

Select the first image:NextOrigWorms

Multiply the first image by:1

Measurement:

Image or measurement?:Image

Select the second image:

Multiply the second image by:1

Measurement:

IdentifyPrimaryObjects:[module_num:7|svn_version:\'Unknown\'|variable_revision_number:10|show_window:True|notes:\x5B\x5D|batch_state:array(\x5B\x5D, dtype=uint8)|enabled:True|wants_pause:False]

Select the input image:PrevOrigWorms

Name the primary objects to be identified:Bubbles

Typical diameter of objects, in pixel units (Min,Max):10,40

Discard objects outside the diameter range?:Yes

Try to merge too small objects with nearby larger objects?:No

Discard objects touching the border of the image?:Yes
Method to distinguish clumped objects:None
Method to draw dividing lines between clumped objects:Intensity
Size of smoothing filter:10
Suppress local maxima that are closer than this minimum allowed distance:7.0
Speed up by using lower-resolution image to find local maxima?:Yes
Name the outline image:PrimaryOutlines
Fill holes in identified objects?:After both thresholding and declumping
Automatically calculate size of smoothing filter for declumping?:Yes
Automatically calculate minimum allowed distance between local maxima?:Yes
Retain outlines of the identified objects?:No
Automatically calculate the threshold using the Otsu method?:Yes
Enter Laplacian of Gaussian threshold:0.5
Automatically calculate the size of objects for the Laplacian of Gaussian filter?:Yes
Enter LoG filter diameter:5.0
Handling of objects if excessive number of objects identified:Continue
Maximum number of objects:500
Threshold setting version:1
Threshold strategy:Global
Thresholding method:Otsu
Select the smoothing method for thresholding:Automatic
Threshold smoothing scale:1.0
Threshold correction factor:2
Lower and upper bounds on threshold:0.001,0.015
Approximate fraction of image covered by objects?:0.01
Manual threshold:0.0
Select the measurement to threshold with:None
Select binary image:None
Masking objects:None
Two-class or three-class thresholding?:Three classes
Minimize the weighted variance or the entropy?:Weighted variance
Assign pixels in the middle intensity class to the foreground or the background?:Foreground
Method to calculate adaptive window size:Image size
Size of adaptive window:10
Use default parameters?:Default
Lower outlier fraction:0.05
Upper outlier fraction:0.05
Averaging method:Mean
Variance method:Standard deviation
# of deviations:2.0

ConvertObjectsToImage:[module_num:8|svn_version:\'Unknown\'|variable_revision_number:1|show_window:False|notes:\x5B\x5D|batch_state:array(\x5B\x5D, dtype=uint8)|enabled:True|wants_pause:False]
Select the input objects:Bubbles
Name the output image:BubbleImage
Select the color format:Binary (black & white)
Select the colormap:Default

Morph:[module_num:9|svn_version:\'Unknown\'|variable_revision_number:4|show_window:True|notes:\x5B\x5D|batch_state:array(\x5B\x5D, dtype=uint8)|enabled:True|wants_pause:False]
Select the input image:BubbleImage
Name the output image:FilledBubbleMask
Select the operation to perform:close
Number of times to repeat operation:Once
Repetition number:2

Diameter:10
Structuring element:Disk
X offset:1.0
Y offset:1.0
Angle:0.0
Width:3.0
Height:3.0
Custom:5,5,1111111111111111111111111
Rescale values from 0 to 1?:Yes

IdentifyPrimaryObjects:[module_num:10|svn_version:\'Unknown\'|variable_revision_number:10|show_window:True|notes:\x5B\x5D|batch_state:array(\x5B\x5D, dtype=uint8)|enabled:True|wants_pause:False]
Select the input image:PrevOrigWorms
Name the primary objects to be identified:Well
Typical diameter of objects, in pixel units (Min,Max):450,550
Discard objects outside the diameter range?:Yes
Try to merge too small objects with nearby larger objects?:No
Discard objects touching the border of the image?:No
Method to distinguish clumped objects:None
Method to draw dividing lines between clumped objects:None
Size of smoothing filter:10
Suppress local maxima that are closer than this minimum allowed distance:5
Speed up by using lower-resolution image to find local maxima?:Yes
Name the outline image:DoNotSave
Fill holes in identified objects?:After both thresholding and declumping
Automatically calculate size of smoothing filter for declumping?:Yes
Automatically calculate minimum allowed distance between local maxima?:Yes
Retain outlines of the identified objects?:No
Automatically calculate the threshold using the Otsu method?:Yes
Enter Laplacian of Gaussian threshold:.5
Automatically calculate the size of objects for the Laplacian of Gaussian filter?:Yes
Enter LoG filter diameter:5
Handling of objects if excessive number of objects identified:Continue
Maximum number of objects:500
Threshold setting version:1
Threshold strategy:Adaptive
Thresholding method:Otsu
Select the smoothing method for thresholding:Automatic
Threshold smoothing scale:10
Threshold correction factor:0.7
Lower and upper bounds on threshold:0,0.01
Approximate fraction of image covered by objects?:0.01
Manual threshold:0.0
Select the measurement to threshold with:None
Select binary image:Otsu Global
Masking objects:From image
Two-class or three-class thresholding?:Three classes
Minimize the weighted variance or the entropy?:Weighted variance
Assign pixels in the middle intensity class to the foreground or the background?:Foreground
Method to calculate adaptive window size:Image size
Size of adaptive window:10
Use default parameters?:Default
Lower outlier fraction:0.05
Upper outlier fraction:0.05
Averaging method:Mean

Variance method:Standard deviation
# of deviations:2.0

FlagImage:[module_num:11|svn_version:\'Unknown\'|variable_revision_number:4|show_window:False|notes:\x5B\x5D|batch_state:array(\x5B\x5D, dtype=uint8)|enabled:True|wants_pause:False]
Hidden:1
Hidden:1
Name the flag\'s category:QC_Flag
Name the flag:Well_Count
How should measurements be linked?:Flag if any fail
Skip image set if flagged?:Yes
Flag is based on:Whole-image measurement
Select the object to be used for flagging:None
Which measurement?:Count_Well
Flag images based on low values?:Yes
Minimum value:1
Flag images based on high values?:No
Maximum value:1.0
Rules file location:Elsewhere...\x7C
Rules file name:rules.txt
Class number:

ConvertObjectsToImage:[module_num:12|svn_version:\'Unknown\'|variable_revision_number:1|show_window:False|notes:\x5B\x5D|batch_state:array(\x5B\x5D, dtype=uint8)|enabled:True|wants_pause:False]
Select the input objects:Well
Name the output image:ExpandedRawWellImage
Select the color format:Binary (black & white)
Select the colormap:Default

Morph:[module_num:13|svn_version:\'Unknown\'|variable_revision_number:4|show_window:True|notes:\x5B\x5D|batch_state:array(\x5B\x5D, dtype=uint8)|enabled:True|wants_pause:False]
Select the input image:ExpandedRawWellImage
Name the output image:ConvexHullWellMask
Select the operation to perform:convex hull
Number of times to repeat operation:Once
Repetition number:2
Diameter:25
Structuring element:Disk
X offset:1.0
Y offset:1.0
Angle:0.0
Width:3.0
Height:3.0
Custom:5,5,1111111111111111111111111
Rescale values from 0 to 1?:Yes

MaskImage:[module_num:14|svn_version:\'Unknown\'|variable_revision_number:3|show_window:False|notes:\x5B\x5D|batch_state:array(\x5B\x5D, dtype=uint8)|enabled:True|wants_pause:False]
Select the input image:PrevOrigWorms
Name the output image:PrevBubbleMaskedOrigWorms
Use objects or an image as a mask?:Image
Select object for mask:ShrunkenWell
Select image for mask:FilledBubbleMask
Invert the mask?:Yes

MaskImage:[module_num:15|svn_version:\'Unknown\'|variable_revision_number:3|show_window:True|notes:\x5B"This time, we\'ll use a different correctio method on the original image. Mask the original image external to the well. CorrectIlluminationCalculate will take the mask into account when computing the illumination function."\x5D|batch_state:array(\x5B\x5D, dtype=uint8)|enabled:True|wants_pause:False]

Select the input image:PrevBubbleMaskedOrigWorms

Name the output image:PrevBubbleWellMaskedOrigWorms

Use objects or an image as a mask?:Image

Select object for mask:ShrunkenWell

Select image for mask:ConvexHullWellMask

Invert the mask?:No

MaskImage:[module_num:16|svn_version:\'Unknown\'|variable_revision_number:3|show_window:False|notes:\x5B\x5D|batch_state:array(\x5B\x5D, dtype=uint8)|enabled:True|wants_pause:False]

Select the input image:NextOrigWorms

Name the output image:NextBubbleMaskedOrigWorms

Use objects or an image as a mask?:Image

Select object for mask:ShrunkenWell

Select image for mask:FilledBubbleMask

Invert the mask?:Yes

MaskImage:[module_num:17|svn_version:\'Unknown\'|variable_revision_number:3|show_window:False|notes:\x5B\x5D|batch_state:array(\x5B\x5D, dtype=uint8)|enabled:True|wants_pause:False]

Select the input image:NextBubbleMaskedOrigWorms

Name the output image:NextBubbleWellMaskedOrigWorms

Use objects or an image as a mask?:Image

Select object for mask:ShrunkenWell

Select image for mask:ConvexHullWellMask

Invert the mask?:No

CorrectIlluminationCalculate:[module_num:18|svn_version:\'Unknown\'|variable_revision_number:2|show_window:True|notes:\x5B"Perform background correction using the convex hull method; see the help for \'Smoothing method\' for more details on how this method works. "\x5D|batch_state:array(\x5B\x5D, dtype=uint8)|enabled:True|wants_pause:False]

Select the input image:PrevBubbleWellMaskedOrigWorms

Name the output image:ConvexHullIllumWorm

Select how the illumination function is calculated:Regular

Dilate objects in the final averaged image?:No

Dilation radius:0

Block size:2

Rescale the illumination function?:Yes

Calculate function for each image individually, or based on all images?:Each

Smoothing method:Convex Hull

Method to calculate smoothing filter size:Object size

Approximate object size:75

Smoothing filter size:10

Retain the averaged image?:No

Name the averaged image:Do not save

Retain the dilated image?:No

Name the dilated image:Do not save

Automatically calculate spline parameters?:Yes

Background mode:auto

Number of spline points:5

Background threshold:2

Image resampling factor:2

Maximum number of iterations:40

Residual value for convergence:0.001

CorrectIlluminationApply:[module_num:19|svn_version:\'Unknown\'|variable_revision_number:3|show_window:False|notes:\x5B\'Apply the illumination function to the original image by division and examine the result. The background is effectively removed from the original image. The corrected image would then need to be inverted using ImageMath for object identification.\'\x5D|batch_state:array(\x5B\x5D, dtype=uint8)|enabled:True|wants_pause:False]

Select the input image:PrevBubbleWellMaskedOrigWorms

Name the output image:PrevConvexHullCorrWorm

Select the illumination function:ConvexHullIllumWorm

Select how the illumination function is applied:Divide

CorrectIlluminationApply:[module_num:20|svn_version:\'Unknown\'|variable_revision_number:3|show_window:False|notes:\x5B\x5D|batch_state:array(\x5B\x5D, dtype=uint8)|enabled:True|wants_pause:False]

Select the input image:NextBubbleWellMaskedOrigWorms

Name the output image:NextConvexHullCorrWorm

Select the illumination function:ConvexHullIllumWorm

Select how the illumination function is applied:Divide

ImageMath:[module_num:21|svn_version:\'Unknown\'|variable_revision_number:4|show_window:True|notes:\x5B\x5D|batch_state:array(\x5B\x5D, dtype=uint8)|enabled:True|wants_pause:False]

Operation:Absolute Difference

Raise the power of the result by:1.0

Multiply the result by:1.0

Add to result:0.0

Set values less than 0 equal to 0?:Yes

Set values greater than 1 equal to 1?:Yes

Ignore the image masks?:No

Name the output image:DiffWorms

Image or measurement?:Image

Select the first image:PrevConvexHullCorrWorm

Multiply the first image by:1.0

Measurement:

Image or measurement?:Image

Select the second image:NextConvexHullCorrWorm

Multiply the second image by:1.0

Measurement:

ImageMath:[module_num:22|svn_version:\'Unknown\'|variable_revision_number:4|show_window:False|notes:\x5B\x5D|batch_state:array(\x5B\x5D, dtype=uint8)|enabled:True|wants_pause:False]

Operation:Invert

Raise the power of the result by:1

Multiply the result by:1

Add to result:0

Set values less than 0 equal to 0?:Yes

Set values greater than 1 equal to 1?:Yes

Ignore the image masks?:No

Name the output image:PrevInvertedConvexHullCorrWorms

Image or measurement?:Image

Select the first image:PrevConvexHullCorrWorm

Multiply the first image by:1

Measurement:

Image or measurement?:Image

Select the second image:

Multiply the second image by:1

Measurement:

ImageMath:[module_num:23|svn_version:\'Unknown\'|variable_revision_number:4|show_window:False|notes:\x5B\x5D|batch_state:array(\x5B\x5D, dtype=uint8)|enabled:True|wants_pause:False]

Operation:Invert

Raise the power of the result by:1

Multiply the result by:1

Add to result:0

Set values less than 0 equal to 0?:Yes

Set values greater than 1 equal to 1?:Yes

Ignore the image masks?:No

Name the output image:NextInvertedConvexHullCorrWorms

Image or measurement?:Image

Select the first image:NextConvexHullCorrWorm

Multiply the first image by:1

Measurement:

Image or measurement?:Image

Select the second image:

Multiply the second image by:1

Measurement:

IdentifyPrimaryObjects:[module_num:24|svn_version:\'Unknown\'|variable_revision_number:10|show_window:True|notes:\x5B"In this step, the goal is to identify all single worms. The \'Typical diameter of objects\' is the diameter of a disc with the same area as the object. Adjust this setting so that small debris and large clusters are excluded. ", \'\', \'If the worms are poorly defined over the image background, adjust the thresholding method. It is preferable to use the same thresholding setting here as during worm untangling since the threshold can affect the size of the worms, making the model less representative of the data set.\'\x5D|batch_state:array(\x5B\x5D, dtype=uint8)|enabled:True|wants_pause:False]

Select the input image:PrevInvertedConvexHullCorrWorms

Name the primary objects to be identified:PrevWormObjects

Typical diameter of objects, in pixel units (Min,Max):40,250

Discard objects outside the diameter range?:Yes

Try to merge too small objects with nearby larger objects?:No

Discard objects touching the border of the image?:No

Method to distinguish clumped objects:None

Method to draw dividing lines between clumped objects:Intensity

Size of smoothing filter:10

Suppress local maxima that are closer than this minimum allowed distance:7

Speed up by using lower-resolution image to find local maxima?:Yes

Name the outline image:PrevWormOutlines

Fill holes in identified objects?:Never

Automatically calculate size of smoothing filter for declumping?:Yes

Automatically calculate minimum allowed distance between local maxima?:Yes

Retain outlines of the identified objects?:No

Automatically calculate the threshold using the Otsu method?:Yes

Enter Laplacian of Gaussian threshold:0.5

Automatically calculate the size of objects for the Laplacian of Gaussian filter?:Yes

Enter LoG filter diameter:5

Handling of objects if excessive number of objects identified:Continue

Maximum number of objects:500

Threshold setting version:1

Threshold strategy:Automatic

Thresholding method:Otsu

Select the smoothing method for thresholding:Automatic

Threshold smoothing scale:1

Threshold correction factor:0.98

Lower and upper bounds on threshold:0.90,1.000000

Approximate fraction of image covered by objects?:0.01

Manual threshold:0.0
Select the measurement to threshold with:None
Select binary image:None
Masking objects:From image
Two-class or three-class thresholding?:Three classes
Minimize the weighted variance or the entropy?:Weighted variance
Assign pixels in the middle intensity class to the foreground or the background?:Background
Method to calculate adaptive window size:Image size
Size of adaptive window:10
Use default parameters?:Default
Lower outlier fraction:0.05
Upper outlier fraction:0.05
Averaging method:Mean
Variance method:Standard deviation
# of deviations:2.0

IdentifyPrimaryObjects:[module_num:25|svn_version:\'Unknown\'|variable_revision_number:10|show_window:True|notes:\x5B\x5D|batch_state:array(\x5B\x5D, dtype=uint8)|enabled:True|wants_pause:False]
Select the input image:NextInvertedConvexHullCorrWorms
Name the primary objects to be identified:NextWormObjects
Typical diameter of objects, in pixel units (Min,Max):40,250
Discard objects outside the diameter range?:Yes
Try to merge too small objects with nearby larger objects?:No
Discard objects touching the border of the image?:No
Method to distinguish clumped objects:None
Method to draw dividing lines between clumped objects:Intensity
Size of smoothing filter:10
Suppress local maxima that are closer than this minimum allowed distance:7
Speed up by using lower-resolution image to find local maxima?:Yes
Name the outline image:NeWormOutlines
Fill holes in identified objects?:Never
Automatically calculate size of smoothing filter for declumping?:Yes
Automatically calculate minimum allowed distance between local maxima?:Yes
Retain outlines of the identified objects?:No
Automatically calculate the threshold using the Otsu method?:Yes
Enter Laplacian of Gaussian threshold:0.5
Automatically calculate the size of objects for the Laplacian of Gaussian filter?:Yes
Enter LoG filter diameter:5
Handling of objects if excessive number of objects identified:Continue
Maximum number of objects:500
Threshold setting version:1
Threshold strategy:Automatic
Thresholding method:Otsu
Select the smoothing method for thresholding:Automatic
Threshold smoothing scale:1
Threshold correction factor:0.98
Lower and upper bounds on threshold:0.90,1.000000
Approximate fraction of image covered by objects?:0.01
Manual threshold:0.0
Select the measurement to threshold with:None
Select binary image:None
Masking objects:From image
Two-class or three-class thresholding?:Three classes
Minimize the weighted variance or the entropy?:Weighted variance
Assign pixels in the middle intensity class to the foreground or the background?:Background

Method to calculate adaptive window size:Image size
Size of adaptive window:10
Use default parameters?:Default
Lower outlier fraction:0.05
Upper outlier fraction:0.05
Averaging method:Mean
Variance method:Standard deviation
# of deviations:2.0

MeasureObjectSizeShape:[module_num:26|svn_version:\'Unknown\'|variable_revision_number:1|show_window:False|notes:\x5B"When measuring size and shape, use the \'OverlappingWorms\' object, which treats even overlapping worms as distinct objects. The Zernike features takes a while to calculate, so we exclude them here."\x5D|batch_state:array(\x5B\x5D, dtype=uint8)|enabled:True|wants_pause:False]
Select objects to measure:PrevWormObjects
Select objects to measure:NextWormObjects
Select objects to measure:Bubbles
Calculate the Zernike features?:No

MeasureObjectIntensity:[module_num:27|svn_version:\'Unknown\'|variable_revision_number:3|show_window:False|notes:\x5B"For intensity measurements, specify the objects to measure and the images from which measurements should be made. When measuring intensities, results may be more accurate if overlapping regions are excluded. Therefore, chose \'NonOverlappingWorms\' as the input object. Select the fluorescent \'Sytox\' image for intensity measurements. For some phenotypes, the bright-field image may also provide informative intensity measurements."\x5D|batch_state:array(\x5B\x5D, dtype=uint8)|enabled:True|wants_pause:False]
Hidden:1
Select an image to measure:PrevInvertedConvexHullCorrWorms
Select objects to measure:PrevWormObjects

MeasureObjectIntensity:[module_num:28|svn_version:\'Unknown\'|variable_revision_number:3|show_window:False|notes:\x5B\x5D|batch_state:array(\x5B\x5D, dtype=uint8)|enabled:True|wants_pause:False]
Hidden:1
Select an image to measure:NextInvertedConvexHullCorrWorms
Select objects to measure:NextWormObjects

FilterObjects:[module_num:29|svn_version:\'Unknown\'|variable_revision_number:7|show_window:True|notes:\x5B\x5D|batch_state:array(\x5B\x5D, dtype=uint8)|enabled:True|wants_pause:False]
Name the output objects:PrevFilteredWorms
Select the object to filter:PrevWormObjects
Select the filtering mode:Measurements
Select the filtering method:Limits
Select the objects that contain the filtered objects:None
Retain outlines of the identified objects?:Yes
Name the outline image:PrevFilteredWormOutlines
Rules file location:Elsewhere...\x7C
Rules file name:rules.txt
Class number:1
Measurement count:5
Additional object count:0
Assign overlapping child to:Both parents
Select the measurement to filter by:AreaShape_Area
Filter using a minimum measurement value?:Yes
Minimum value:2100
Filter using a maximum measurement value?:No
Maximum value:1.0
Select the measurement to filter by:Intensity_IntegratedIntensityEdge_PrevInvertedConvexHullCorrWorms
Filter using a minimum measurement value?:Yes

Minimum value:2100
Filter using a maximum measurement value?:No
Maximum value:1.0
Select the measurement to filter by:AreaShape_EulerNumber
Filter using a minimum measurement value?:Yes
Minimum value:-6.0
Filter using a maximum measurement value?:Yes
Maximum value:1.0
Select the measurement to filter by:Intensity_MADIntensity_PrevInvertedConvexHullCorrWorms
Filter using a minimum measurement value?:No
Minimum value:0.01
Filter using a maximum measurement value?:Yes
Maximum value:0.0085
Select the measurement to filter by:Location_MassDisplacement_PrevInvertedConvexHullCorrWorms
Filter using a minimum measurement value?:No
Minimum value:0.1
Filter using a maximum measurement value?:Yes
Maximum value:0.41

FilterObjects:[module_num:30|svn_version:\'Unknown\'|variable_revision_number:7|show_window:True|notes:\x5B\x5D|batch_state:array(\x5B\x5D, dtype=uint8)|enabled:True|wants_pause:False]
Name the output objects:NextFilteredWorms
Select the object to filter:NextWormObjects
Select the filtering mode:Measurements
Select the filtering method:Limits
Select the objects that contain the filtered objects:None
Retain outlines of the identified objects?:Yes
Name the outline image:NextFilteredWormOutlines
Rules file location:Elsewhere...\x7C
Rules file name:rules.txt
Class number:1
Measurement count:5
Additional object count:0
Assign overlapping child to:Both parents
Select the measurement to filter by:AreaShape_Area
Filter using a minimum measurement value?:Yes
Minimum value:2100
Filter using a maximum measurement value?:No
Maximum value:1.0
Select the measurement to filter by:Intensity_IntegratedIntensityEdge_NextInvertedConvexHullCorrWorms
Filter using a minimum measurement value?:Yes
Minimum value:2100
Filter using a maximum measurement value?:No
Maximum value:1.0
Select the measurement to filter by:AreaShape_EulerNumber
Filter using a minimum measurement value?:Yes
Minimum value:-6.0
Filter using a maximum measurement value?:Yes
Maximum value:1.0
Select the measurement to filter by:Intensity_MADIntensity_NextInvertedConvexHullCorrWorms
Filter using a minimum measurement value?:No
Minimum value:0.01
Filter using a maximum measurement value?:Yes
Maximum value:0.0085
Select the measurement to filter by:Intensity_MassDisplacement_NextInvertedConvexHullCorrWorms

Filter using a minimum measurement value?:No
Minimum value:0.0
Filter using a maximum measurement value?:Yes
Maximum value:0.41

FlagImage:[module_num:31|svn_version:\'Unknown\'|variable_revision_number:4|show_window:True|notes:\x5B\x5D|batch_state:array(\x5B\x5D, dtype=uint8)|enabled:True|wants_pause:False]
Hidden:1
Hidden:2
Name the flag\'s category:QC_Flag
Name the flag:WormObjects_Number
How should measurements be linked?:Flag if any fail
Skip image set if flagged?:Yes
Flag is based on:Measurements for all objects in each image
Select the object to be used for flagging:PrevFilteredWorms
Which measurement?:Number_Object_Number
Flag images based on low values?:Yes
Minimum value:1
Flag images based on high values?:Yes
Maximum value:1
Rules file location:Elsewhere...\x7C
Rules file name:rules.txt
Class number:
Flag is based on:Measurements for all objects in each image
Select the object to be used for flagging:NextFilteredWorms
Which measurement?:Number_Object_Number
Flag images based on low values?:Yes
Minimum value:1
Flag images based on high values?:Yes
Maximum value:1.0
Rules file location:Elsewhere...\x7C
Rules file name:rules.txt
Class number:

MeasureObjectSizeShape:[module_num:32|svn_version:\'Unknown\'|variable_revision_number:1|show_window:True|notes:\x5B\x5D|batch_state:array(\x5B\x5D, dtype=uint8)|enabled:True|wants_pause:True]
Select objects to measure:PrevFilteredWorms
Select objects to measure:NextFilteredWorms
Calculate the Zernike features?:Yes

MeasureImageIntensity:[module_num:33|svn_version:\'Unknown\'|variable_revision_number:2|show_window:True|notes:\x5B\x5D|batch_state:array(\x5B\x5D, dtype=uint8)|enabled:True|wants_pause:False]
Select the image to measure:DiffWorms
Measure the intensity only from areas enclosed by objects?:No
Select the input objects:None
Select the image to measure:PrevInvertedConvexHullCorrWorms
Measure the intensity only from areas enclosed by objects?:Yes
Select the input objects:PrevFilteredWorms

MeasureGranularity:[module_num:34|svn_version:\'Unknown\'|variable_revision_number:3|show_window:True|notes:\x5B\x5D|batch_state:array(\x5B\x5D, dtype=uint8)|enabled:True|wants_pause:False]
Image count:1
Object count:1
Select an image to measure:PrevInvertedConvexHullCorrWorms
Subsampling factor for granularity measurements:0.25

Subsampling factor for background reduction:0.25
Radius of structuring element:10
Range of the granular spectrum:16
Select objects to measure:PrevFilteredWorms

MeasureObjectIntensityDistribution:[module_num:35|svn_version:\'Unknown\'|variable_revision_number:5|show_window:True|notes:\x5B\x5D|batch_state:array(\x5B\x5D, dtype=uint8)|enabled:True|wants_pause:False]
Hidden:1
Hidden:1
Hidden:1
Hidden:0
Calculate intensity Zernikes?:None
Maximum zernike moment:9
Select an image to measure:PrevInvertedConvexHullCorrWorms
Select objects to measure:PrevFilteredWorms
Object to use as center?:These objects
Select objects to use as centers:None
Scale the bins?:Yes
Number of bins:4
Maximum radius:100

MeasureTexture:[module_num:36|svn_version:\'Unknown\'|variable_revision_number:4|show_window:True|notes:\x5B\x5D|batch_state:array(\x5B\x5D, dtype=uint8)|enabled:True|wants_pause:False]
Hidden:1
Hidden:1
Hidden:1
Select an image to measure:PrevInvertedConvexHullCorrWorms
Select objects to measure:PrevFilteredWorms
Texture scale to measure:3
Angles to measure:Horizontal,Vertical,Diagonal,Anti-diagonal
Measure Gabor features?:Yes
Number of angles to compute for Gabor:4
Measure images or objects?:Both

MeasureCorrelation:[module_num:37|svn_version:\'Unknown\'|variable_revision_number:3|show_window:True|notes:\x5B\x5D|batch_state:array(\x5B\x5D, dtype=uint8)|enabled:True|wants_pause:False]
Hidden:2
Hidden:2
Select an image to measure:PrevInvertedWorms
Select an image to measure:NextInvertedWorms
Set threshold as percentage of maximum intensity for the images:15.0
Select where to measure correlation:Within objects
Select an object to measure:PrevFilteredWorms
Select an object to measure:NextFilteredWorms

CalculateImageOverlap:[module_num:38|svn_version:\'Unknown\'|variable_revision_number:4|show_window:True|notes:\x5B\x5D|batch_state:array(\x5B\x5D, dtype=uint8)|enabled:True|wants_pause:False]
Compare segmented objects, or foreground/background?:Segmented objects
Select the image to be used as the ground truth basis for calculating the amount of overlap:None
Select the image to be used to test for overlap:None
Select the objects to be used as the ground truth basis for calculating the amount of overlap:PrevFilteredWorms
Select the objects to be tested for overlap against the ground truth:NextFilteredWorms
Calculate earth mover\'s distance?:No
Maximum # of points:250
Point selection method:K Means

Maximum distance:250
Penalize missing pixels:No

OverlayOutlines:[module_num:39|svn_version:\'Unknown\'|variable_revision_number:3|show_window:True|notes:\x5B\'Optional\x3A This module is only for visual examination of the untangling result.\'\x5D|batch_state:array(\x5B\x5D, dtype=uint8)|enabled:True|wants_pause:False]
Display outlines on a blank image?:No
Select image on which to display outlines:PrevOrigWorms
Name the output image:OrigOverlay
Outline display mode:Color
Select method to determine brightness of outlines:Max of image
Width of outlines:1
Select outlines to display:PrevFilteredWormOutlines
Select outline color:Red
Load outlines from an image or objects?:Objects
Select objects to display:PrevFilteredWorms
Select outlines to display:None
Select outline color:Green
Load outlines from an image or objects?:Objects
Select objects to display:NextFilteredWorms

ExportToSpreadsheet:[module_num:40|svn_version:\'Unknown\'|variable_revision_number:11|show_window:True|notes:\x5B\'This module exports measurements to a spreadsheet. You may chose to export all measurements, or limit measurements to those of interest, as shown below. The created csv files include measurements on a per-worm basis, and can be opened in a text editor or any spreadsheet program such as Excel.\'\x5D|batch_state:array(\x5B\x5D, dtype=uint8)|enabled:True|wants_pause:True]
Select the column delimiter:Comma (",")
Add image metadata columns to your object data file?:Yes
Limit output to a size that is allowed in Excel?:No
Select the measurements to export:Yes
Calculate the per-image mean values for object measurements?:No
Calculate the per-image median values for object measurements?:No
Calculate the per-image standard deviation values for object measurements?:No
Output file location:Default Output Folder\x7C\\\\\\\\\\\\\\\iodine-cifs\\\\\\\\imaging_analysis\\\\\\\\2014_09_22_Schistosoma_Floriano_Silva_Oswaldo_Cruz_Foundation\\\\\\\\2014_09_26\\\\\\\\output\\\\\\\\2014-10-02_compare
Create a GenePattern GCT file?:No
Select source of sample row name:Metadata
Select the image to use as the identifier:None
Select the metadata to use as the identifier:None
Export all measurement types?:No

:Image\x7CQC_Flag_WormObjects_Number,Image\x7CQC_Flag_Well_Count,Image\x7CTexture_InfoMeas2_PrevInvertedConvexHullCorrWorms_3_90,Image\x7CTexture_InfoMeas2_PrevInvertedConvexHullCorrWorms_3_45,Image\x7CTexture_InfoMeas2_PrevInvertedConvexHullCorrWorms_3_135,Image\x7CTexture_InfoMeas2_PrevInvertedConvexHullCorrWorms_3_0,Image\x7CTexture_DifferenceVariance_PrevInvertedConvexHullCorrWorms_3_90,Image\x7CTexture_DifferenceVariance_PrevInvertedConvexHullCorrWorms_3_45,Image\x7CTexture_DifferenceVariance_PrevInvertedConvexHullCorrWorms_3_135,Image\x7CTexture_DifferenceVariance_PrevInvertedConvexHullCorrWorms_3_0,Image\x7CTexture_SumVariance_PrevInvertedConvexHullCorrWorms_3_0,Image\x7CTexture_SumVariance_PrevInvertedConvexHullCorrWorms_3_45,Image\x7CTexture_SumVariance_PrevInvertedConvexHullCorrWorms_3_135,Image\x7CTexture_SumVariance_PrevInvertedConvexHullCorrWorms_3_90,Image\x7CTexture_Gabor_PrevInvertedConvexHullCorrWorms_3,Image\x7CTexture_DifferenceEntropy_PrevInvertedConvexHullCorrWorms_3_90,Image\x7CTexture_DifferenceEntropy_PrevInvertedConvexHullCorrWorms_3_45,Image\x7CTexture_DifferenceEntropy_PrevInvertedConvexHullCorrWorms_3_135,Image\x7CTexture_DifferenceEntropy_PrevInvertedConvexHullCorrWorms_3_0,Image\x7CTexture_AngularSecondMoment_PrevInvertedConvexHullCorrWorms_3_0,Image\x7CTexture_AngularSecondMoment_PrevInvertedConvexHullCorrWorms_3_45,Image\x7CTexture_AngularSecondMoment_PrevInvertedConvexHullCorrWorms_3_135,Image\x7CTexture_AngularSecondMoment_PrevInvertedConvexHullCorrWorms_3_90,Image\x7CTexture_Entropy_PrevInvertedConvexHullCorrWorms_3_90,Image\x7CTexture_Entropy_PrevInvertedConvexHullCorrWorms_3

_45,Image\x7CTexture_Entropy_PrevInvertedConvexHullCorrWorms_3_135,Image\x7CTexture_Entropy_PrevInvertedConvexHullCorr
Worms_3_0,Image\x7CTexture_Correlation_PrevInvertedConvexHullCorrWorms_3_0,Image\x7CTexture_Correlation_PrevInvertedCon
vexHullCorrWorms_3_45,Image\x7CTexture_Correlation_PrevInvertedConvexHullCorrWorms_3_135,Image\x7CTexture_Correlation_Pr
evInvertedConvexHullCorrWorms_3_90,Image\x7CTexture_SumAverage_PrevInvertedConvexHullCorrWorms_3_90,Image\x7CTexture
_SumAverage_PrevInvertedConvexHullCorrWorms_3_45,Image\x7CTexture_SumAverage_PrevInvertedConvexHullCorrWorms_3_135,
Image\x7CTexture_SumAverage_PrevInvertedConvexHullCorrWorms_3_0,Image\x7CTexture_Variance_PrevInvertedConvexHullCorrW
orms_3_90,Image\x7CTexture_Variance_PrevInvertedConvexHullCorrWorms_3_45,Image\x7CTexture_Variance_PrevInvertedConvex
HullCorrWorms_3_135,Image\x7CTexture_Variance_PrevInvertedConvexHullCorrWorms_3_0,Image\x7CTexture_SumEntropy_PrevInv
ertedConvexHullCorrWorms_3_0,Image\x7CTexture_SumEntropy_PrevInvertedConvexHullCorrWorms_3_45,Image\x7CTexture_SumE
ntropy_PrevInvertedConvexHullCorrWorms_3_135,Image\x7CTexture_SumEntropy_PrevInvertedConvexHullCorrWorms_3_90,Image\x
7CTexture_InverseDifferenceMoment_PrevInvertedConvexHullCorrWorms_3_90,Image\x7CTexture_InverseDifferenceMoment_PrevInv
ertedConvexHullCorrWorms_3_45,Image\x7CTexture_InverseDifferenceMoment_PrevInvertedConvexHullCorrWorms_3_135,Image\x7
CTexture_InverseDifferenceMoment_PrevInvertedConvexHullCorrWorms_3_0,Image\x7CTexture_Contrast_PrevInvertedConvexHullCo
rrWorms_3_90,Image\x7CTexture_Contrast_PrevInvertedConvexHullCorrWorms_3_45,Image\x7CTexture_Contrast_PrevInvertedConv
exHullCorrWorms_3_135,Image\x7CTexture_Contrast_PrevInvertedConvexHullCorrWorms_3_0,Image\x7CTexture_InfoMeas1_PrevInv
ertedConvexHullCorrWorms_3_90,Image\x7CTexture_InfoMeas1_PrevInvertedConvexHullCorrWorms_3_45,Image\x7CTexture_InfoMe
as1_PrevInvertedConvexHullCorrWorms_3_135,Image\x7CTexture_InfoMeas1_PrevInvertedConvexHullCorrWorms_3_0,Image\x7CFil
eName_NextOrigWorms,Image\x7CFileName_PrevOrigWorms,Image\x7CIntensity_MinIntensity_PrevInvertedConvexHullCorrWorms_P
revFilteredWorms,Image\x7CIntensity_StdIntensity_PrevInvertedConvexHullCorrWorms_PrevFilteredWorms,Image\x7CIntensity_TotalI
ntensity_DiffWorms,Image\x7CIntensity_TotalIntensity_PrevInvertedConvexHullCorrWorms_PrevFilteredWorms,Image\x7CIntensity_Tot
alArea_PrevInvertedConvexHullCorrWorms_PrevFilteredWorms,Image\x7CIntensity_PercentMaximal_PrevInvertedConvexHullCorrWor
ms_PrevFilteredWorms,Image\x7CIntensity_MeanIntensity_DiffWorms,Image\x7CIntensity_MeanIntensity_PrevInvertedConvexHullCorr
Worms_PrevFilteredWorms,Image\x7CIntensity_MADIntensity_PrevInvertedConvexHullCorrWorms_PrevFilteredWorms,Image\x7CInte
nsity_UpperQuartileIntensity_PrevInvertedConvexHullCorrWorms_PrevFilteredWorms,Image\x7CIntensity_MaxIntensity_PrevInvertedC
onvexHullCorrWorms_PrevFilteredWorms,Image\x7CIntensity_MedianIntensity_DiffWorms,Image\x7CIntensity_MedianIntensity_PrevIn
vertedConvexHullCorrWorms_PrevFilteredWorms,Image\x7CIntensity_LowerQuartileIntensity_PrevInvertedConvexHullCorrWorms_Pre
vFilteredWorms,Image\x7CPathName_NextOrigWorms,Image\x7CPathName_PrevOrigWorms,Image\x7COverlap_TrueNegRate_PrevF
ilteredWorms_NextFilteredWorms,Image\x7COverlap_Recall_PrevFilteredWorms_NextFilteredWorms,Image\x7COverlap_TruePosRate
_PrevFilteredWorms_NextFilteredWorms,Image\x7COverlap_Precision_PrevFilteredWorms_NextFilteredWorms,Image\x7COverlap_Ra
ndIndex_PrevFilteredWorms_NextFilteredWorms,Image\x7COverlap_Ffactor_PrevFilteredWorms_NextFilteredWorms,Image\x7COverla
p_AdjustedRandIndex_PrevFilteredWorms_NextFilteredWorms,Image\x7COverlap_FalseNegRate_PrevFilteredWorms_NextFilteredWo
rms,Image\x7COverlap_FalsePosRate_PrevFilteredWorms_NextFilteredWorms,Image\x7CGranularity_11_PrevInvertedConvexHullCorr
Worms,Image\x7CGranularity_10_PrevInvertedConvexHullCorrWorms,Image\x7CGranularity_13_PrevInvertedConvexHullCorrWorms,I
mage\x7CGranularity_12_PrevInvertedConvexHullCorrWorms,Image\x7CGranularity_15_PrevInvertedConvexHullCorrWorms,Image\x7
CGranularity_14_PrevInvertedConvexHullCorrWorms,Image\x7CGranularity_16_PrevInvertedConvexHullCorrWorms,Image\x7CGranul
arity_1_PrevInvertedConvexHullCorrWorms,Image\x7CGranularity_3_PrevInvertedConvexHullCorrWorms,Image\x7CGranularity_2_Pre
vInvertedConvexHullCorrWorms,Image\x7CGranularity_5_PrevInvertedConvexHullCorrWorms,Image\x7CGranularity_4_PrevInvertedC
onvexHullCorrWorms,Image\x7CGranularity_7_PrevInvertedConvexHullCorrWorms,Image\x7CGranularity_6_PrevInvertedConvexHullC
orrWorms,Image\x7CGranularity_9_PrevInvertedConvexHullCorrWorms,Image\x7CGranularity_8_PrevInvertedConvexHullCorrWorms,I
mage\x7CMetadata_Plate,Image\x7CMetadata_Column,Image\x7CMetadata_Gender,Image\x7CMetadata_Time_h,Image\x7CMetadat
a_Well,Image\x7CMetadata_Compound,Image\x7CMetadata_Compound_micM,Image\x7CMetadata_Row,PrevFilteredWorms\x7CPare
nt_PrevWormObjects,PrevFilteredWorms\x7CCorrelation_Correlation_PrevInvertedWorms_NextInvertedWorms,PrevFilteredWorms\x7C
Granularity_11_PrevInvertedConvexHullCorrWorms,PrevFilteredWorms\x7CGranularity_10_PrevInvertedConvexHullCorrWorms,PrevFil
teredWorms\x7CGranularity_13_PrevInvertedConvexHullCorrWorms,PrevFilteredWorms\x7CGranularity_12_PrevInvertedConvexHullC
orrWorms,PrevFilteredWorms\x7CGranularity_15_PrevInvertedConvexHullCorrWorms,PrevFilteredWorms\x7CGranularity_14_PrevInve
rtedConvexHullCorrWorms,PrevFilteredWorms\x7CGranularity_16_PrevInvertedConvexHullCorrWorms,PrevFilteredWorms\x7CGranula
rity_1_PrevInvertedConvexHullCorrWorms,PrevFilteredWorms\x7CGranularity_3_PrevInvertedConvexHullCorrWorms,PrevFilteredWor
ms\x7CGranularity_2_PrevInvertedConvexHullCorrWorms,PrevFilteredWorms\x7CGranularity_5_PrevInvertedConvexHullCorrWorms,Pr
evFilteredWorms\x7CGranularity_4_PrevInvertedConvexHullCorrWorms,PrevFilteredWorms\x7CGranularity_7_PrevInvertedConvexHull
CorrWorms,PrevFilteredWorms\x7CGranularity_6_PrevInvertedConvexHullCorrWorms,PrevFilteredWorms\x7CGranularity_9_PrevInvert
edConvexHullCorrWorms,PrevFilteredWorms\x7CGranularity_8_PrevInvertedConvexHullCorrWorms,PrevFilteredWorms\x7CTexture_Di
fferenceEntropy_PrevInvertedConvexHullCorrWorms_3_90,PrevFilteredWorms\x7CTexture_DifferenceEntropy_PrevInvertedConvexHull
CorrWorms_3_45,PrevFilteredWorms\x7CTexture_DifferenceEntropy_PrevInvertedConvexHullCorrWorms_3_135,PrevFilteredWorms\x
7CTexture_DifferenceEntropy_PrevInvertedConvexHullCorrWorms_3_0,PrevFilteredWorms\x7CTexture_DifferenceVariance_PrevInvert
edConvexHullCorrWorms_3_90,PrevFilteredWorms\x7CTexture_DifferenceVariance_PrevInvertedConvexHullCorrWorms_3_45,PrevFilt

eredWorms\x7CTexture_DifferenceVariance_PrevInvertedConvexHullCorrWorms_3_135,PrevFilteredWorms\x7CTexture_DifferenceVariance_PrevInvertedConvexHullCorrWorms_3_0,PrevFilteredWorms\x7CTexture_InfoMeas1_PrevInvertedConvexHullCorrWorms_3_90,PrevFilteredWorms\x7CTexture_InfoMeas1_PrevInvertedConvexHullCorrWorms_3_45,PrevFilteredWorms\x7CTexture_InfoMeas1_PrevInvertedConvexHullCorrWorms_3_135,PrevFilteredWorms\x7CTexture_InfoMeas1_PrevInvertedConvexHullCorrWorms_3_0,PrevFilteredWorms\x7CTexture_SumVariance_PrevInvertedConvexHullCorrWorms_3_90,PrevFilteredWorms\x7CTexture_SumVariance_PrevInvertedConvexHullCorrWorms_3_45,PrevFilteredWorms\x7CTexture_SumVariance_PrevInvertedConvexHullCorrWorms_3_135,PrevFilteredWorms\x7CTexture_SumVariance_PrevInvertedConvexHullCorrWorms_3_0,PrevFilteredWorms\x7CTexture_Gabor_PrevInvertedConvexHullCorrWorms_3,PrevFilteredWorms\x7CTexture_AngularSecondMoment_PrevInvertedConvexHullCorrWorms_3_0,PrevFilteredWorms\x7CTexture_AngularSecondMoment_PrevInvertedConvexHullCorrWorms_3_45,PrevFilteredWorms\x7CTexture_AngularSecondMoment_PrevInvertedConvexHullCorrWorms_3_135,PrevFilteredWorms\x7CTexture_AngularSecondMoment_PrevInvertedConvexHullCorrWorms_3_90,PrevFilteredWorms\x7CTexture_Entropy_PrevInvertedConvexHullCorrWorms_3_90,PrevFilteredWorms\x7CTexture_Entropy_PrevInvertedConvexHullCorrWorms_3_45,PrevFilteredWorms\x7CTexture_Entropy_PrevInvertedConvexHullCorrWorms_3_135,PrevFilteredWorms\x7CTexture_Entropy_PrevInvertedConvexHullCorrWorms_3_0,PrevFilteredWorms\x7CTexture_Correlation_PrevInvertedConvexHullCorrWorms_3_0,PrevFilteredWorms\x7CTexture_Correlation_PrevInvertedConvexHullCorrWorms_3_45,PrevFilteredWorms\x7CTexture_Correlation_PrevInvertedConvexHullCorrWorms_3_135,PrevFilteredWorms\x7CTexture_Correlation_PrevInvertedConvexHullCorrWorms_3_90,PrevFilteredWorms\x7CTexture_SumAverage_PrevInvertedConvexHullCorrWorms_3_0,PrevFilteredWorms\x7CTexture_SumAverage_PrevInvertedConvexHullCorrWorms_3_45,PrevFilteredWorms\x7CTexture_SumAverage_PrevInvertedConvexHullCorrWorms_3_135,PrevFilteredWorms\x7CTexture_SumAverage_PrevInvertedConvexHullCorrWorms_3_90,PrevFilteredWorms\x7CTexture_Variance_PrevInvertedConvexHullCorrWorms_3_90,PrevFilteredWorms\x7CTexture_Variance_PrevInvertedConvexHullCorrWorms_3_45,PrevFilteredWorms\x7CTexture_Variance_PrevInvertedConvexHullCorrWorms_3_135,PrevFilteredWorms\x7CTexture_Variance_PrevInvertedConvexHullCorrWorms_3_0,PrevFilteredWorms\x7CTexture_SumEntropy_PrevInvertedConvexHullCorrWorms_3_90,PrevFilteredWorms\x7CTexture_SumEntropy_PrevInvertedConvexHullCorrWorms_3_45,PrevFilteredWorms\x7CTexture_SumEntropy_PrevInvertedConvexHullCorrWorms_3_135,PrevFilteredWorms\x7CTexture_SumEntropy_PrevInvertedConvexHullCorrWorms_3_0,PrevFilteredWorms\x7CTexture_InverseDifferenceMoment_PrevInvertedConvexHullCorrWorms_3_90,PrevFilteredWorms\x7CTexture_InverseDifferenceMoment_PrevInvertedConvexHullCorrWorms_3_45,PrevFilteredWorms\x7CTexture_InverseDifferenceMoment_PrevInvertedConvexHullCorrWorms_3_135,PrevFilteredWorms\x7CTexture_InverseDifferenceMoment_PrevInvertedConvexHullCorrWorms_3_0,PrevFilteredWorms\x7CTexture_Contrast_PrevInvertedConvexHullCorrWorms_3_0,PrevFilteredWorms\x7CTexture_Contrast_PrevInvertedConvexHullCorrWorms_3_45,PrevFilteredWorms\x7CTexture_Contrast_PrevInvertedConvexHullCorrWorms_3_135,PrevFilteredWorms\x7CTexture_Contrast_PrevInvertedConvexHullCorrWorms_3_90,PrevFilteredWorms\x7CTexture_InfoMeas2_PrevInvertedConvexHullCorrWorms_3_0,PrevFilteredWorms\x7CTexture_InfoMeas2_PrevInvertedConvexHullCorrWorms_3_45,PrevFilteredWorms\x7CTexture_InfoMeas2_PrevInvertedConvexHullCorrWorms_3_135,PrevFilteredWorms\x7CTexture_InfoMeas2_PrevInvertedConvexHullCorrWorms_3_90,PrevFilteredWorms\x7CNumber_Object_Number,PrevFilteredWorms\x7CLocation_Center_Y,PrevFilteredWorms\x7CLocation_Center_X,PrevFilteredWorms\x7CAreaShape_Perimeter,PrevFilteredWorms\x7CAreaShape_FormFactor,PrevFilteredWorms\x7CAreaShape_MinorAxisLength,PrevFilteredWorms\x7CAreaShape_Center_Y,PrevFilteredWorms\x7CAreaShape_Center_X,PrevFilteredWorms\x7CAreaShape_Area,PrevFilteredWorms\x7CAreaShape_MinFeretDiameter,PrevFilteredWorms\x7CAreaShape_Solidity,PrevFilteredWorms\x7CAreaShape_MaxFeretDiameter,PrevFilteredWorms\x7CAreaShape_Zernike_1_1,PrevFilteredWorms\x7CAreaShape_Zernike_0_0,PrevFilteredWorms\x7CAreaShape_Zernike_3_1,PrevFilteredWorms\x7CAreaShape_Zernike_3_3,PrevFilteredWorms\x7CAreaShape_Zernike_2_0,PrevFilteredWorms\x7CAreaShape_Zernike_2_2,PrevFilteredWorms\x7CAreaShape_Zernike_5_1,PrevFilteredWorms\x7CAreaShape_Zernike_5_3,PrevFilteredWorms\x7CAreaShape_Zernike_5_5,PrevFilteredWorms\x7CAreaShape_Zernike_4_0,PrevFilteredWorms\x7CAreaShape_Zernike_4_2,PrevFilteredWorms\x7CAreaShape_Zernike_4_4,PrevFilteredWorms\x7CAreaShape_Zernike_7_1,PrevFilteredWorms\x7CAreaShape_Zernike_7_3,PrevFilteredWorms\x7CAreaShape_Zernike_7_5,PrevFilteredWorms\x7CAreaShape_Zernike_7_7,PrevFilteredWorms\x7CAreaShape_Zernike_6_0,PrevFilteredWorms\x7CAreaShape_Zernike_6_6,PrevFilteredWorms\x7CAreaShape_Zernike_6_2,PrevFilteredWorms\x7CAreaShape_Zernike_6_4,PrevFilteredWorms\x7CAreaShape_Zernike_9_1,PrevFilteredWorms\x7CAreaShape_Zernike_9_3,PrevFilteredWorms\x7CAreaShape_Zernike_9_5,PrevFilteredWorms\x7CAreaShape_Zernike_9_7,PrevFilteredWorms\x7CAreaShape_Zernike_9_9,PrevFilteredWorms\x7CAreaShape_Zernike_8_0,PrevFilteredWorms\x7CAreaShape_Zernike_8_2,PrevFilteredWorms\x7CAreaShape_Zernike_8_4,PrevFilteredWorms\x7CAreaShape_Zernike_8_6,PrevFilteredWorms\x7CAreaShape_Zernike_8_8,PrevFilteredWorms\x7CAreaShape_Eccentricity,PrevFilteredWorms\x7CAreaShape_Compactness,PrevFilteredWorms\x7CAreaShape_Extent,PrevFilteredWorms\x7CAreaShape_Orientation,PrevFilteredWorms\x7CAreaShape_MedianRadius,PrevFilteredWorms\x7CAreaShape_MaximumRadius,PrevFilteredWorms\x7CAreaShape_EulerNumber,PrevFilteredWorms\x7CAreaShape_MajorAxisLength,PrevFilteredWorms\x7CAreaShape_MeanRadius,PrevFilteredWorms\x7CRadialDistribution_FracAtD_PrevInvertedConvexHullCorrWorms_2of4,PrevFilteredWorms\x7CRadialDistribution_FracAtD_PrevInvertedConvexHullCorrWorms_3of4,PrevFilteredWorms\x7CRadialDistribution_FracAtD_PrevInvertedConvexHullCorrWorms_1of4,PrevFilteredWorms\x7CRadialDistribution_FracAtD_PrevInvertedConvexHullCorrWorms_4of4,PrevFilteredWorms\x7CRadialDistribution_RadialCV_PrevInvertedConvexHullCorrWorms_2of4,PrevFilteredWorms\x7CRadialDistribution_RadialCV_PrevInvertedConvexHullCorrWorms_3of4,PrevFilteredWorms\x7CRadialDistribution_RadialCV_PrevInvertedConvexHullCorrWorms_1of4,PrevFilteredWorms\x7CRadialDistribution_RadialCV_Prev

InvertedConvexHullCorrWorms_4of4,PrevFilteredWorms\x7CRadialDistribution_MeanFrac_PrevInvertedConvexHullCorrWorms_2of4,PrevFilteredWorms\x7CRadialDistribution_MeanFrac_PrevInvertedConvexHullCorrWorms_3of4,PrevFilteredWorms\x7CRadialDistribution_MeanFrac_PrevInvertedConvexHullCorrWorms_4of4,PrevFilteredWorms\x7CRadialDistribution_MeanFrac_PrevInvertedConvexHullCorrWorms_1of4,Bubbles\x7CAreaShape_Perimeter,Bubbles\x7CAreaShape_FormFactor,Bubbles\x7CAreaShape_Solidity,Bubbles\x7CAreaShape_Orientation,Bubbles\x7CAreaShape_Area,Bubbles\x7CAreaShape_MinFeretDiameter,Bubbles\x7CAreaShape_MajorAxisLength,Bubbles\x7CAreaShape_MaxFeretDiameter,Bubbles\x7CAreaShape_EulerNumber,Bubbles\x7CAreaShape_Eccentricity,Bubbles\x7CAreaShape_Compactness,Bubbles\x7CAreaShape_Extent,Bubbles\x7CAreaShape_MedianRadius,Bubbles\x7CAreaShape_MaximumRadius,Bubbles\x7CAreaShape_MeanRadius,Bubbles\x7CAreaShape_MinorAxisLength,Bubbles\x7CAreaShape_Center_Y,Bubbles\x7CAreaShape_Center_X

Representation of Nan/Inf:NaN

Add a prefix to file names?:No

Filename prefix:MyExpt_

Overwrite existing files without warning?:Yes

Data to export:Image

Combine these object measurements with those of the previous object?:No

File name:DATA.csv

Use the object name for the file name?:Yes

Data to export:PrevFilteredWorms

Combine these object measurements with those of the previous object?:Yes

File name:DATA.csv

Use the object name for the file name?:Yes

ExportToDatabase:[module_num:41|svn_version:\'Unknown\'|variable_revision_number:27|show_window:True|notes:\x5B\x5D|batch_state:array(\x5B\x5D, dtype=uint8)|enabled:True|wants_pause:False]

Database type:MySQL / CSV

Database name:Schisto_HCS-DB

Add a prefix to Tabela names?:Yes

Tabela prefix:MyExpt_

SQL file prefix:SQL_

Output file location:Default Output Folder\x7C

Create a CellProfiler Analyst properties file?:Yes

Database host:jarvis

Username:bioensaios

Password:

Name the SQLite database file:DefaultDB.db

Calculate the per-image mean values of object measurements?:No

Calculate the per-image median values of object measurements?:No

Calculate the per-image standard deviation values of object measurements?:No

Calculate the per-well mean values of object measurements?:Yes

Calculate the per-well median values of object measurements?:Yes

Calculate the per-well standard deviation values of object measurements?:Yes

Export measurements for all objects to the database?:Select...

Select the objects:PrevFilteredWorms

Maximum # of characters in a column name:64

Create one Tabela per object, a single object Tabela or a single object view?:Single object Tabela

Enter an image url prepend if you plan to access your files via http:

Write image thumbnails directly to the database?:No

Select the images for which you want to save thumbnails:

Auto-scale thumbnail pixel intensities?:Yes

Select the plate type:96

Select the plate metadata:Plate

Select the well metadata:Well

Include information for all images, using default values?:Yes

Properties image group count:1

Properties group field count:5

Properties filter field count:0
Workspace measurement count:1
Experiment name:MyExpt
Which objects should be used for locations?:PrevFilteredWorms
Enter a phenotype class Tabela name if using the classifier tool:Drug_X
Export object relationships?:Yes
Overwrite without warning?:Never
Access CPA images via URL?:No
Select the classification type:Object
Select an image to include:None
Use the image name for the display?:Yes
Image name:Channel1
Channel color:red
Do you want to add group fields?:Yes
Enter the name of the group:Well
Enter the per-image columns which define the group, separated by commas:Image_Metadata_Plate, Image_Metadata_Well
Enter the name of the group:Gender
Enter the per-image columns which define the group, separated by commas:Image_Metadata_Gender
Enter the name of the group:Time
Enter the per-image columns which define the group, separated by commas:Image_Metadata_Time
Enter the name of the group:Drug_Concentration
Enter the per-image columns which define the group, separated by commas:Image_Metadata_conc
Enter the name of the group:Group
Enter the per-image columns which define the group, separated by commas:Image_Metadata_Group
Do you want to add filter fields?:No
Automatically create a filter for each plate?:No
Create a CellProfiler Analyst workspace file?:Yes
Select the measurement display tool:ScatterPlot
Type of measurement to plot on the X-axis:Image
Enter the object name:None
Select the X-axis measurement:Metadata_conc
Select the X-axis index:ImageNumber
Type of measurement to plot on the Y-axis:Image
Enter the object name:None
Select the Y-axis measurement:Overlap_AdjustedRandIndex_PrevFilteredWorms_NextFilteredWorms
Select the Y-axis index:ImageNumber

7.2 Anexo 2: *Script* para a implementação das funções (em ambiente R) requeridas para predição de coeficientes de pertencimento para novas observações a partir dos resultados de agrupamento gerados pelos dados de treinamento.

```
## funçao objetiva renovada:
# (com ela minimizada teremos a predicao de um elemento extra no grupo e das memberships)
```

```
#Esta eh uma funcao para gerar a funcao objetiva com F maiusculo a ser minimizada
f.ob.new <-
  function(Train, A.train, obs.test){
    #
    K = A.train$k.crisp
    M = A.train$memb.exp
    #
    D.train = fDiss(Train %>%
                      select(starts_with("Comp")),
                    method = "mahalanobis", center=F,scale=F)
    #
    D.obs.train = fDiss(Xr = Train %>%
                          select(starts_with("Comp")),
                        X2 = obs.test%>%
                          select(starts_with("Comp")),
                        method = "mahalanobis", center=F,scale=F)
    #
    C1 = rep(NA,K)
    C2 = rep(NA,K)
    C3 = rep(NA,K)
    for(v in 1:K){
      C1[v] = sum( ( ((A.train$membership[,v])^M) * D.train )%*%(A.train$membership[,v])^M )
      C2[v] = 2*sum( ((A.train$membership[,v])^M) * D.obs.train )
      C3[v] = sum( (A.train$membership[,v])^M )
    }
    # onde substituicao x^2 = u
    # F = sum( (C1 + C2*(X^(2*M))) / (2*( C3 + (X^(2*M)))) )
    #
    #
    expres <-
      parse(text = paste('F.obj.new <- function(x,m) { return( sum( (' ,
                         paste0('c(', paste(C1, collapse = ","),')' ),
                         ' + ',
                         paste0('c(', paste(C2, collapse = ","),')' ),
                         '*(x^(2*m)))',
                         ' / ',
                         '(2*( ',
                         paste0('c(', paste(C3, collapse = ","),')' ),
                         ' + (x^(2*m))))',
                         ' )',
                         ') }', sep=''))
```

```
    return(expres)

  }

#Esta eh uma funcao para gerar o gradiente da funcao objetiva com F maiusculo a ser minimizada
grad.f.ob.new <-
  function(Train, A.train, obs.test){
    #
    K = A.train$k.crisp
    M = A.train$memb.exp
    #
    D.train = fDiss(Train %>%
                      select(starts_with("Comp")),
                    method = "mahalanobis", center=F,scale=F)
    #
    D.obs.train = fDiss(Xr = Train %>%
                          select(starts_with("Comp")),
                        X2 = obs.test%>%
                          select(starts_with("Comp")),
                        method = "mahalanobis", center=F,scale=F)
    #
    C1 = rep(NA,K)
    C2 = rep(NA,K)
    C3 = rep(NA,K)
    for(v in 1:K){
      C1[v] = sum( ( ((A.train$membership[,v])^M) * D.train )%*%(A.train$membership[,v])^M )
      C2[v] = 2*sum( ((A.train$membership[,v])^M) * D.obs.train )
      C3[v] = sum( (A.train$membership[,v])^M )
    }
    # onde substituicao x^2 = u
    # gradF[v] =    ( m*x[v]^(2*m-1) )*(C2[v]*C3[v] - C1[v]) / (C3[v] + x^(2*m))^2

    #
    expres <-
      parse(text = paste('gradF.obj.new <- function(x,m) { return(    (m*(x^(2*m-1)) )*(' ,
                         paste0('c(', paste(C2, collapse = ","),')' ),
                         ' * ',
                         paste0('c(', paste(C3, collapse = ","),')' ),
                         ' - ',
                         paste0('c(', paste(C1, collapse = ","),')' ),
                         ' )',
```

```
                                    ' / ',
                                    ' (( ',
                                    paste0('c(', paste(C3, collapse = ","),')' ),
                                    ' + (x^(2*m)))^2 )',
                                    ' ) ',
                                    ' }', sep=''))
    return(expres)

  }


# funcao valor inicial
inicial <- function(Train, A.train, Test){
  d.maha = fDiss(Xr = Train%>%
                      select(starts_with("Comp")),
                    X2 = Test%>%
                      select(starts_with("Comp")),
                    method = "mahalanobis", center=F,scale=F)
  indice = apply(d.maha, MARGIN = 2, FUN = function(x) which(x==min(x)))
  chute = sqrt(A.train$membership[indice,])
  return(chute)
}

#
# funcao minimizacao para predizer memberships dos dados Test nos dados Train (ela tem como prerequisito a existencia de objetos no ambiente referentes aos dados de treinamento Train e aos resultados do agrupamento de interesse A.train.
minimizacao <-
  function(obs, Train, A.train){
    eval(f.ob.new(Train = Train,
                  A.train =   A.train,
                  obs.test = obs))
    eval(grad.f.ob.new(Train = Train,
                       A.train =   A.train,
                       obs.test = obs))

    # assegurando a funcao
        constrOptim.nl2 <- purrr::possibly(alabama::constrOptim.nl, otherwise = NA, quiet = F)
    #
    x.otimo <-
      constrOptim.nl2(
```

```
            par = valor.inicial.x[obs$indice,],
            fn = function(x){F.obj.new(x,m=m)},
            gr = function(x){gradF.obj.new(x,m=m)},
            heq = function(x){sum(x^2) - 1},
            control.outer = list(trace=FALSE)
          )


    if( !anyNA(x.otimo) ){
       x_otimo <- x.otimo$par
       membership.predicit <- (x_otimo)^2
    } else{ membership.predicit <- rep(NA,k) }

    return(membership.predicit)
  }
```

7.3 Anexo 3: Dicionário de rótulos originais e substitutos

A relação a seguir foi construída usando a letra inicial do rótulo original (para fazer referência ao módulo do qual a variável é derivada) e um rótulo númerico (sem qualquer significado). Os rótulos originais descrevem, respectivamente: os módulos de origem das medidas; os tipos de medida; as imagens/objetos usados como *input* para a geração da medida; e demais parâmetros usados no cálculo (usando a nomenclatura do *script* do *pipeline* de processamento de imagens, Anexo 1).

Rótulo Original	Rótulo curto
Overlap_AdjustedRandIndex_PrevFilteredWorms_NextFilteredWorms	O_007
Overlap_FalseNegRate_PrevFilteredWorms_NextFilteredWorms	O_008
Overlap_FalsePosRate_PrevFilteredWorms_NextFilteredWorms	O_009
Overlap_Ffactor_PrevFilteredWorms_NextFilteredWorms	O_010
Overlap_Precision_PrevFilteredWorms_NextFilteredWorms	O_011
Overlap_RandIndex_PrevFilteredWorms_NextFilteredWorms	O_012
Overlap_Recall_PrevFilteredWorms_NextFilteredWorms	O_013
Overlap_TrueNegRate_PrevFilteredWorms_NextFilteredWorms	O_014
Overlap_TruePosRate_PrevFilteredWorms_NextFilteredWorms	O_015
Intensity_LowerQuartileIntensity_PrevInvertedConvexHullCorrWorms_PrevFilteredWorms	I_019

Intensity_MADIntensity_PrevInvertedConvexHullCorrWorms_PrevFilteredWorms	I_020
Intensity_MaxIntensity_PrevInvertedConvexHullCorrWorms_PrevFilteredWorms	I_021
Intensity_MeanIntensity_DiffWorms	I_022
Intensity_MeanIntensity_PrevInvertedConvexHullCorrWorms_PrevFilteredWorms	I_023
Intensity_MedianIntensity_DiffWorms	I_024
Intensity_MedianIntensity_PrevInvertedConvexHullCorrWorms_PrevFilteredWorms	I_025
Intensity_MinIntensity_PrevInvertedConvexHullCorrWorms_PrevFilteredWorms	I_026
Intensity_PercentMaximal_PrevInvertedConvexHullCorrWorms_PrevFilteredWorms	I_027
Intensity_StdIntensity_PrevInvertedConvexHullCorrWorms_PrevFilteredWorms	I_028
Intensity_TotalArea_PrevInvertedConvexHullCorrWorms_PrevFilteredWorms	I_029
Intensity_TotalIntensity_DiffWorms	I_030
Intensity_TotalIntensity_PrevInvertedConvexHullCorrWorms_PrevFilteredWorms	I_031
Intensity_UpperQuartileIntensity_PrevInvertedConvexHullCorrWorms_PrevFilteredWorms	I_032
Granularity_1_PrevInvertedConvexHullCorrWorms	G_033
Granularity_2_PrevInvertedConvexHullCorrWorms	G_034
Granularity_3_PrevInvertedConvexHullCorrWorms	G_035
Granularity_4_PrevInvertedConvexHullCorrWorms	G_036
Granularity_5_PrevInvertedConvexHullCorrWorms	G_037
Granularity_6_PrevInvertedConvexHullCorrWorms	G_038
Granularity_7_PrevInvertedConvexHullCorrWorms	G_039
Granularity_8_PrevInvertedConvexHullCorrWorms	G_040
Granularity_9_PrevInvertedConvexHullCorrWorms	G_041
Granularity_10_PrevInvertedConvexHullCorrWorms	G_042
Granularity_11_PrevInvertedConvexHullCorrWorms	G_043
Granularity_12_PrevInvertedConvexHullCorrWorms	G_044
Granularity_13_PrevInvertedConvexHullCorrWorms	G_045
Granularity_14_PrevInvertedConvexHullCorrWorms	G_046
Granularity_15_PrevInvertedConvexHullCorrWorms	G_047
Granularity_16_PrevInvertedConvexHullCorrWorms	G_048
Texture_AngularSecondMoment_PrevInvertedConvexHullCorrWorms_3_0	T_049
Texture_AngularSecondMoment_PrevInvertedConvexHullCorrWorms_3_135	T_050
Texture_AngularSecondMoment_PrevInvertedConvexHullCorrWorms_3_45	T_051
Texture_AngularSecondMoment_PrevInvertedConvexHullCorrWorms_3_90	T_052
Texture_Contrast_PrevInvertedConvexHullCorrWorms_3_0	T_053
Texture_Contrast_PrevInvertedConvexHullCorrWorms_3_135	T_054
Texture_Contrast_PrevInvertedConvexHullCorrWorms_3_45	T_055
Texture_Contrast_PrevInvertedConvexHullCorrWorms_3_90	T_056

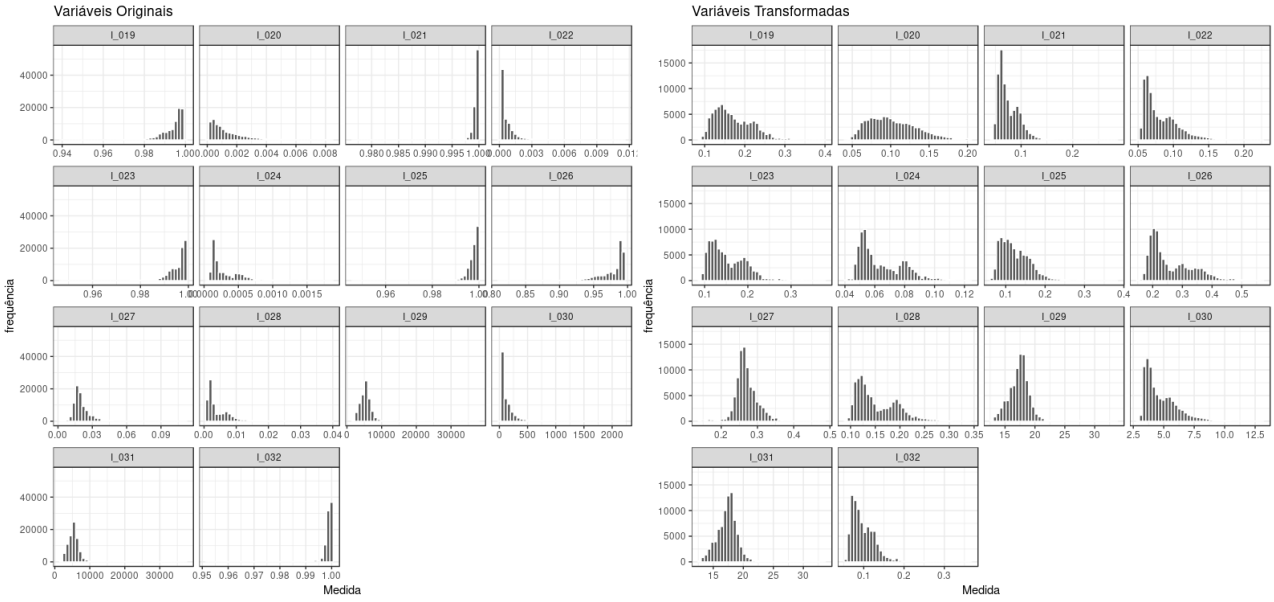
Texture_Correlation_PrevInvertedConvexHullCorrWorms_3_0	T_057
Texture_Correlation_PrevInvertedConvexHullCorrWorms_3_135	T_058
Texture_Correlation_PrevInvertedConvexHullCorrWorms_3_45	T_059
Texture_Correlation_PrevInvertedConvexHullCorrWorms_3_90	T_060
Texture_DifferenceEntropy_PrevInvertedConvexHullCorrWorms_3_0	T_061
Texture_DifferenceEntropy_PrevInvertedConvexHullCorrWorms_3_135	T_062
Texture_DifferenceEntropy_PrevInvertedConvexHullCorrWorms_3_45	T_063
Texture_DifferenceEntropy_PrevInvertedConvexHullCorrWorms_3_90	T_064
Texture_DifferenceVariance_PrevInvertedConvexHullCorrWorms_3_0	T_065
Texture_DifferenceVariance_PrevInvertedConvexHullCorrWorms_3_135	T_066
Texture_DifferenceVariance_PrevInvertedConvexHullCorrWorms_3_45	T_067
Texture_DifferenceVariance_PrevInvertedConvexHullCorrWorms_3_90	T_068
Texture_Entropy_PrevInvertedConvexHullCorrWorms_3_0	T_069
Texture_Entropy_PrevInvertedConvexHullCorrWorms_3_135	T_070
Texture_Entropy_PrevInvertedConvexHullCorrWorms_3_45	T_071
Texture_Entropy_PrevInvertedConvexHullCorrWorms_3_90	T_072
Texture_Gabor_PrevInvertedConvexHullCorrWorms_3	T_073
Texture_InfoMeas1_PrevInvertedConvexHullCorrWorms_3_0	T_074
Texture_InfoMeas1_PrevInvertedConvexHullCorrWorms_3_135	T_075
Texture_InfoMeas1_PrevInvertedConvexHullCorrWorms_3_45	T_076
Texture_InfoMeas1_PrevInvertedConvexHullCorrWorms_3_90	T_077
Texture_InfoMeas2_PrevInvertedConvexHullCorrWorms_3_0	T_078
Texture_InfoMeas2_PrevInvertedConvexHullCorrWorms_3_135	T_079
Texture_InfoMeas2_PrevInvertedConvexHullCorrWorms_3_45	T_080
Texture_InfoMeas2_PrevInvertedConvexHullCorrWorms_3_90	T_081
Texture_InverseDifferenceMoment_PrevInvertedConvexHullCorrWorms_3_0	T_082
Texture_InverseDifferenceMoment_PrevInvertedConvexHullCorrWorms_3_135	T_083
Texture_InverseDifferenceMoment_PrevInvertedConvexHullCorrWorms_3_45	T_084
Texture_InverseDifferenceMoment_PrevInvertedConvexHullCorrWorms_3_90	T_085
Texture_SumAverage_PrevInvertedConvexHullCorrWorms_3_0	T_086
Texture_SumAverage_PrevInvertedConvexHullCorrWorms_3_135	T_087
Texture_SumAverage_PrevInvertedConvexHullCorrWorms_3_45	T_088
Texture_SumAverage_PrevInvertedConvexHullCorrWorms_3_90	T_089
Texture_SumEntropy_PrevInvertedConvexHullCorrWorms_3_0	T_090
Texture_SumEntropy_PrevInvertedConvexHullCorrWorms_3_135	T_091
Texture_SumEntropy_PrevInvertedConvexHullCorrWorms_3_45	T_092
Texture_SumEntropy_PrevInvertedConvexHullCorrWorms_3_90	T_093

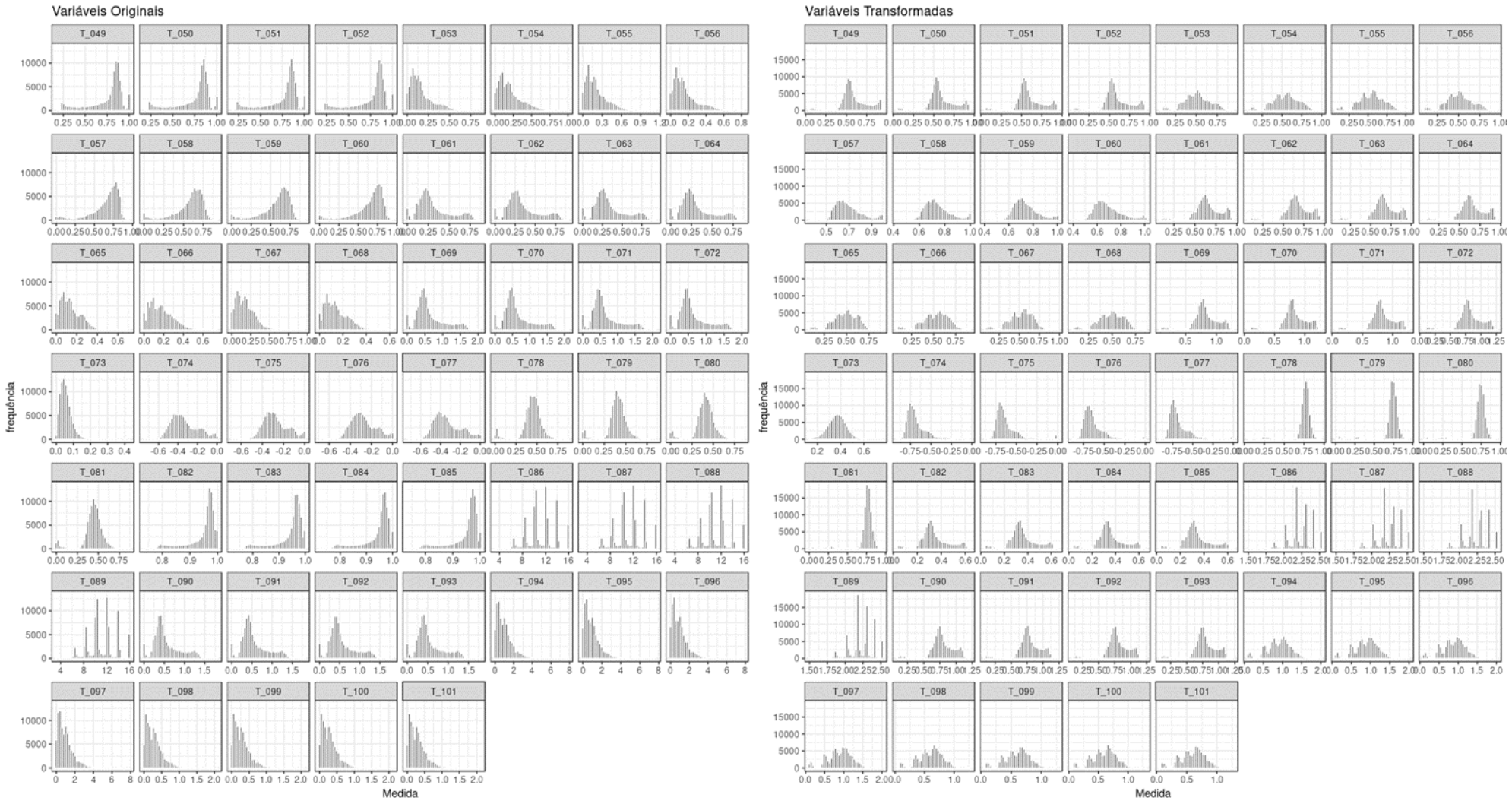
Texture_SumVariance_PrevInvertedConvexHullCorrWorms_3_0	T_094
Texture_SumVariance_PrevInvertedConvexHullCorrWorms_3_135	T_095
Texture_SumVariance_PrevInvertedConvexHullCorrWorms_3_45	T_096
Texture_SumVariance_PrevInvertedConvexHullCorrWorms_3_90	T_097
Texture_Variance_PrevInvertedConvexHullCorrWorms_3_0	T_098
Texture_Variance_PrevInvertedConvexHullCorrWorms_3_135	T_099
Texture_Variance_PrevInvertedConvexHullCorrWorms_3_45	T_100
Texture_Variance_PrevInvertedConvexHullCorrWorms_3_90	T_101

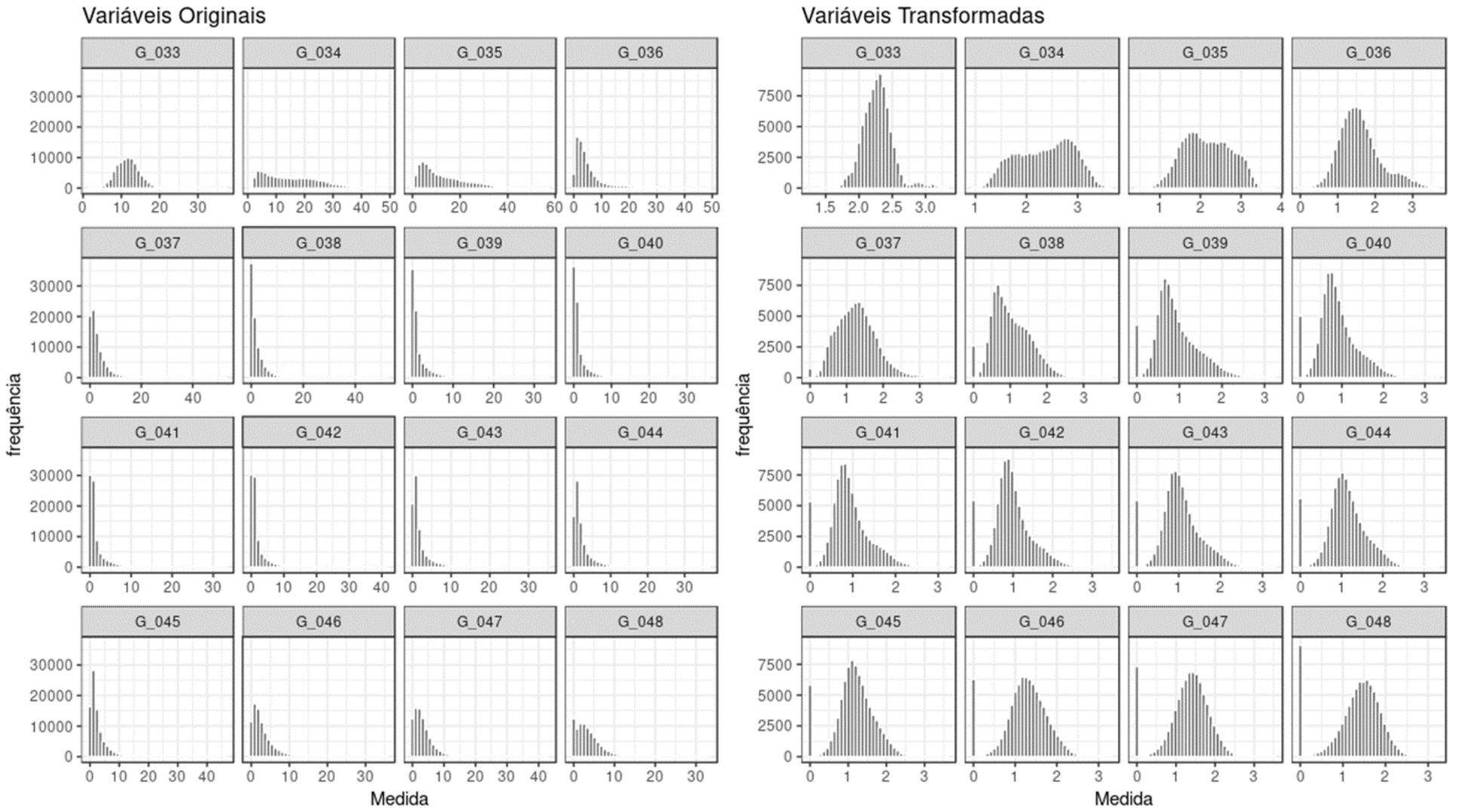
7.4 Anexo 4: Histogramas das variáveis originais e transformadas

Em cada histograma, a faixa dos valores observados é indicada pelo intervalo coberto por cada gráfico. A presença de valores extremos não pode ser vista no histograma pela sua raridade, mas a faixa de valores coberta pelo histograma indica a existência ou não destes.

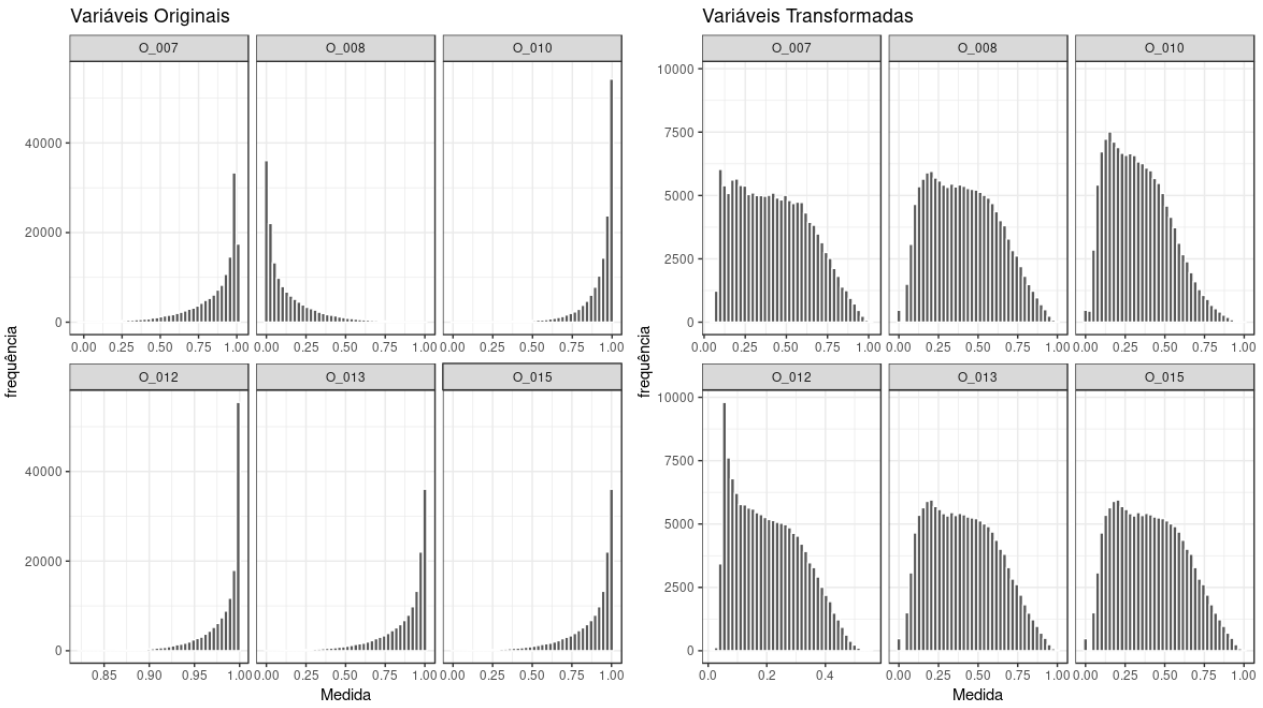
7.4.1 - *Gráficos para os dados de vermes fêmeas*

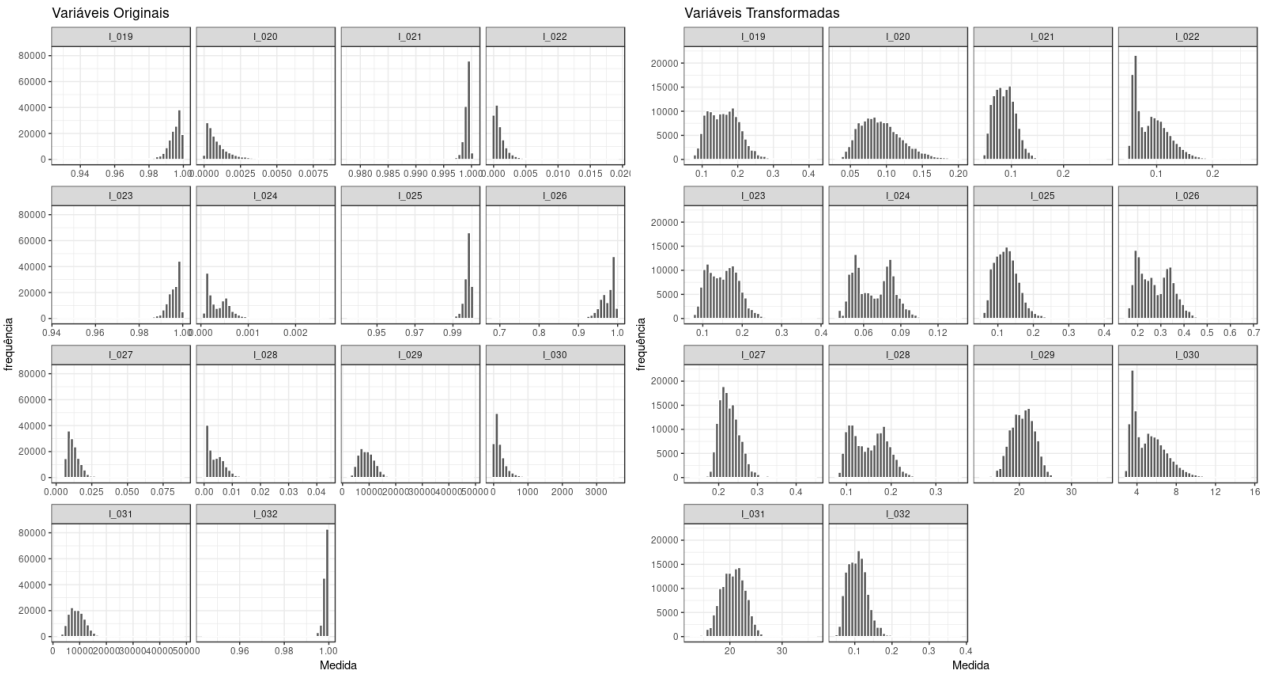


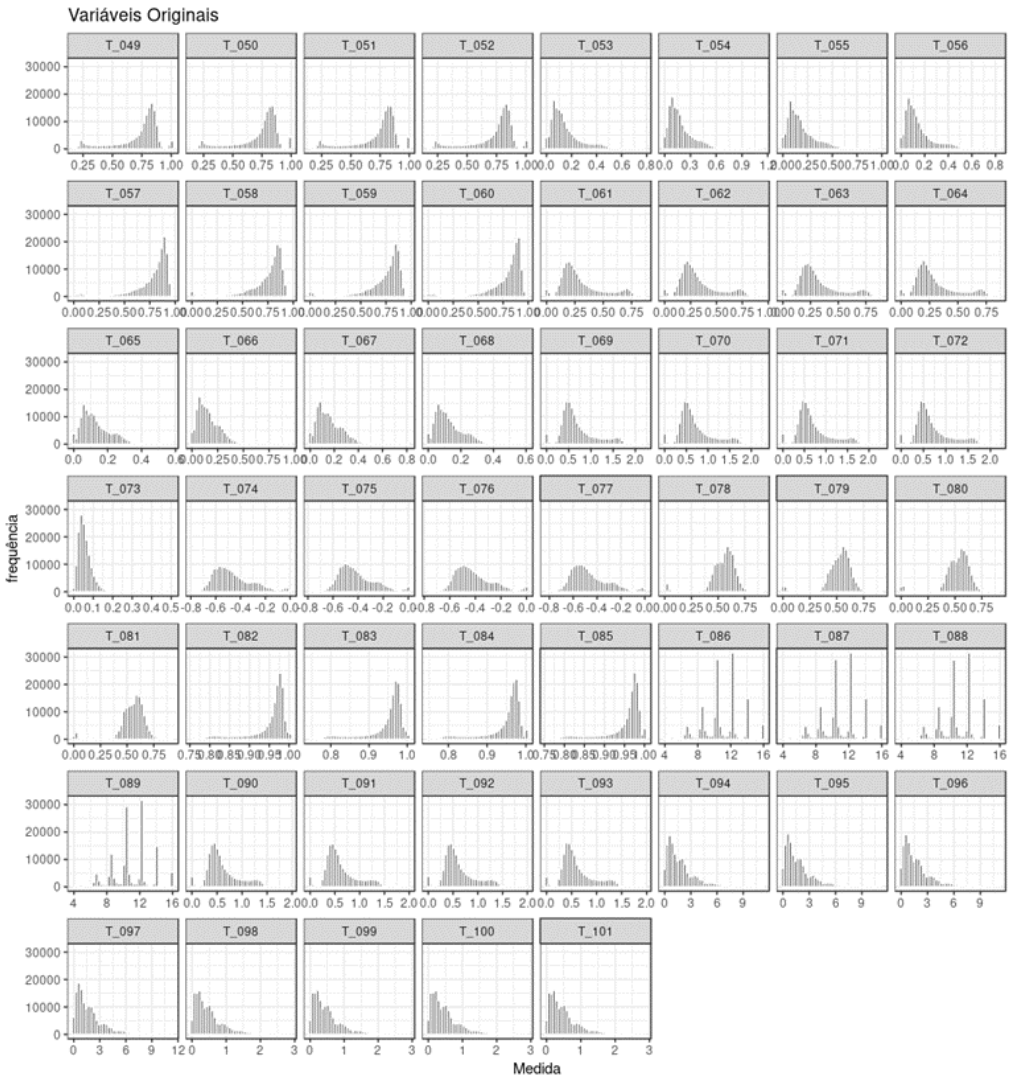

Variáveis Originais
Variáveis Transformadas
frequência
Medida
T_049
T_050
T_051
T_052
T_053
T_054
T_055
T_056
T_057
T_058
T_059
T_060
T_061
T_062
T_063
T_064
T_065
T_066
T_067
T_068
T_069
T_070
T_071
T_072
T_073
T_074
T_075
T_076
T_077
T_078
T_079
T_080
T_081
T_082
T_083
T_084
T_085
T_086
T_087
T_088
T_089
T_090
T_091
T_092
T_093
T_094
T_095
T_096
T_097
T_098
T_099
T_100
T_101

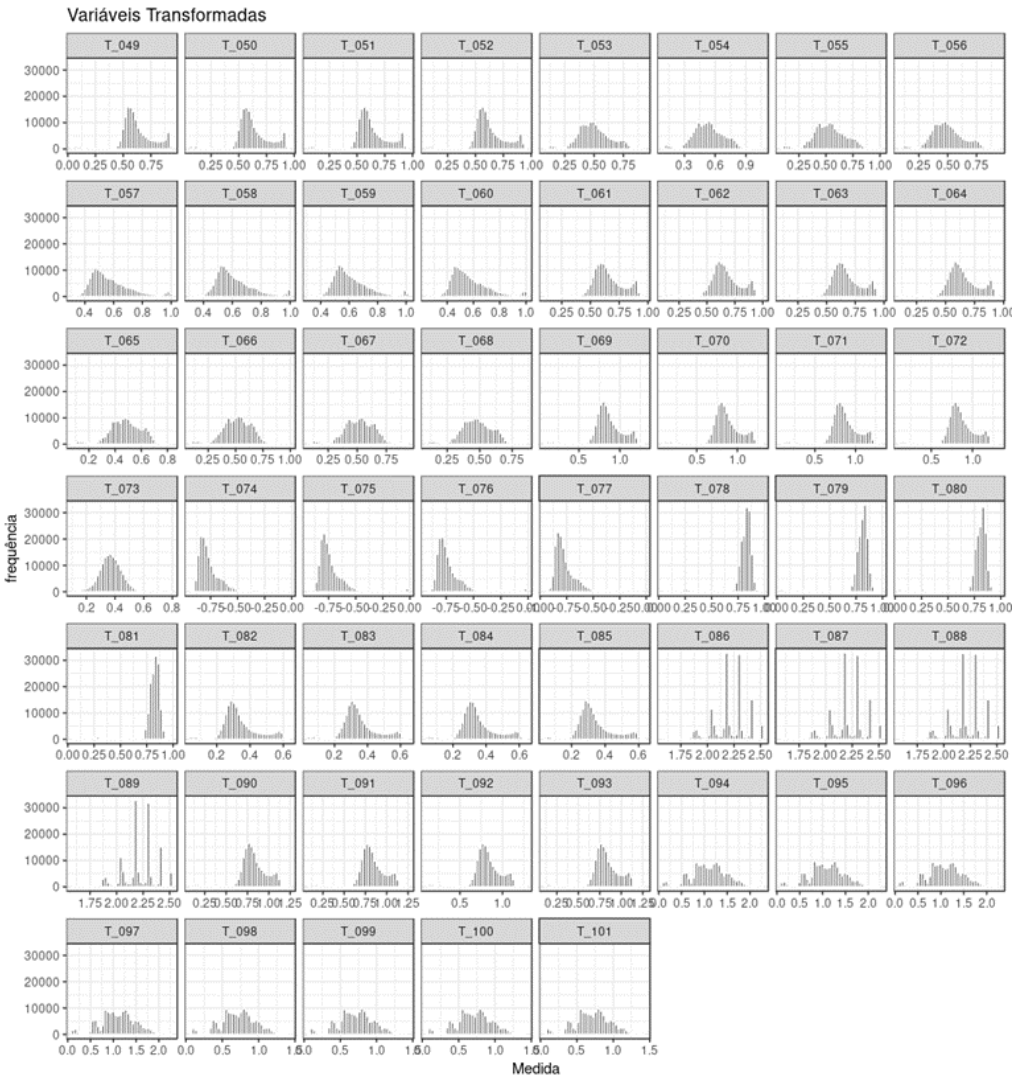

Variáveis Originais
Variáveis Transformadas
G_033
G_034
G_035
G_036
G_037
G_038
G_039
G_040
G_041
G_042
G_043
G_044
G_045
G_046
G_047
G_048
frequência
Medida

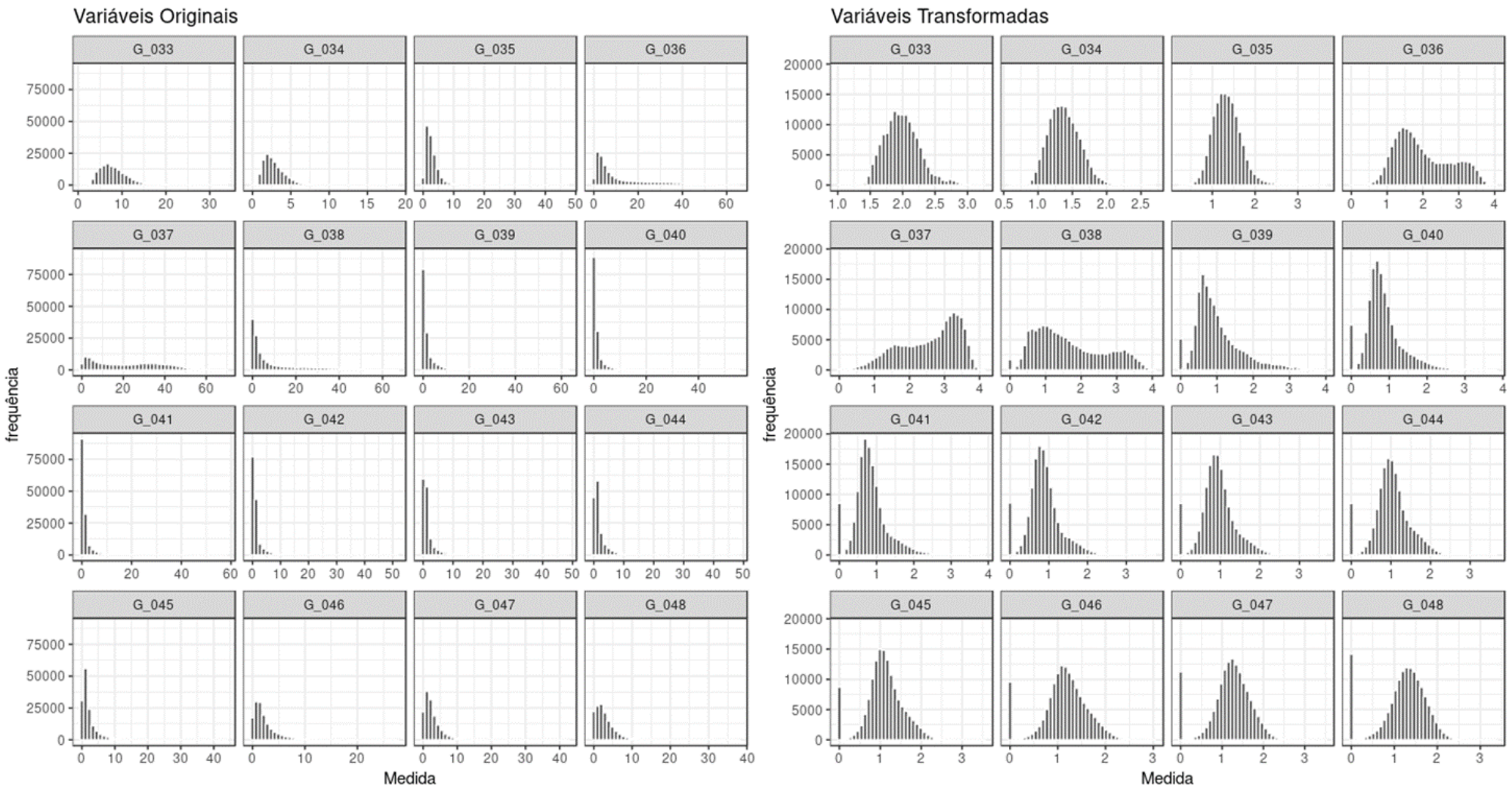
7.4.2 - ***Gráficos para os dados de vermes machos***



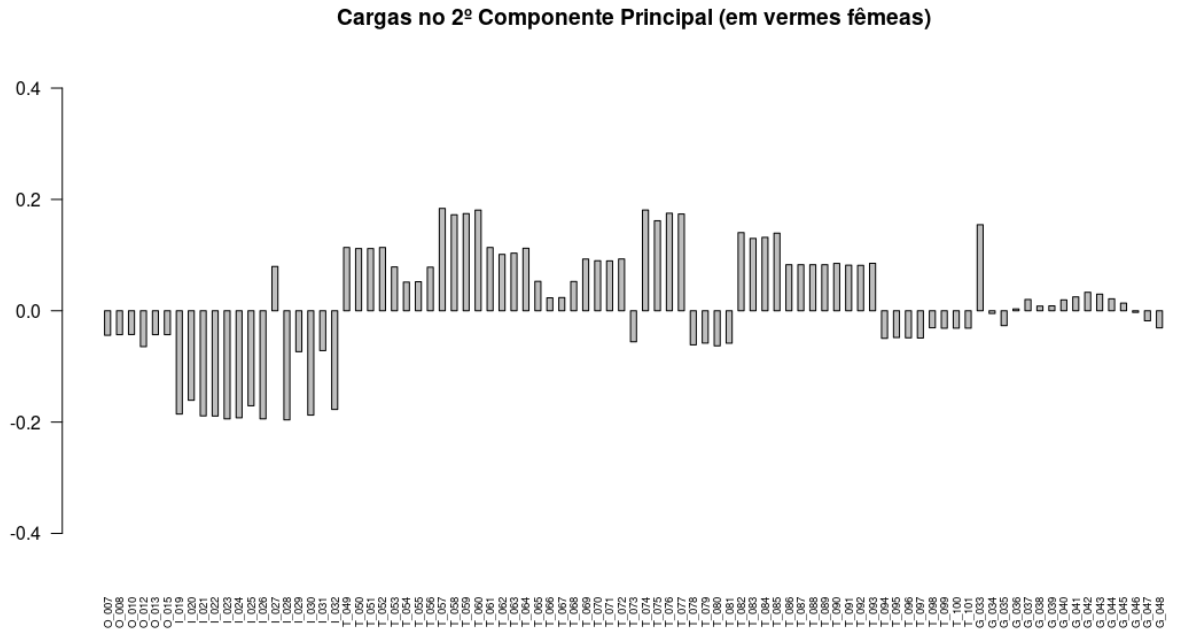


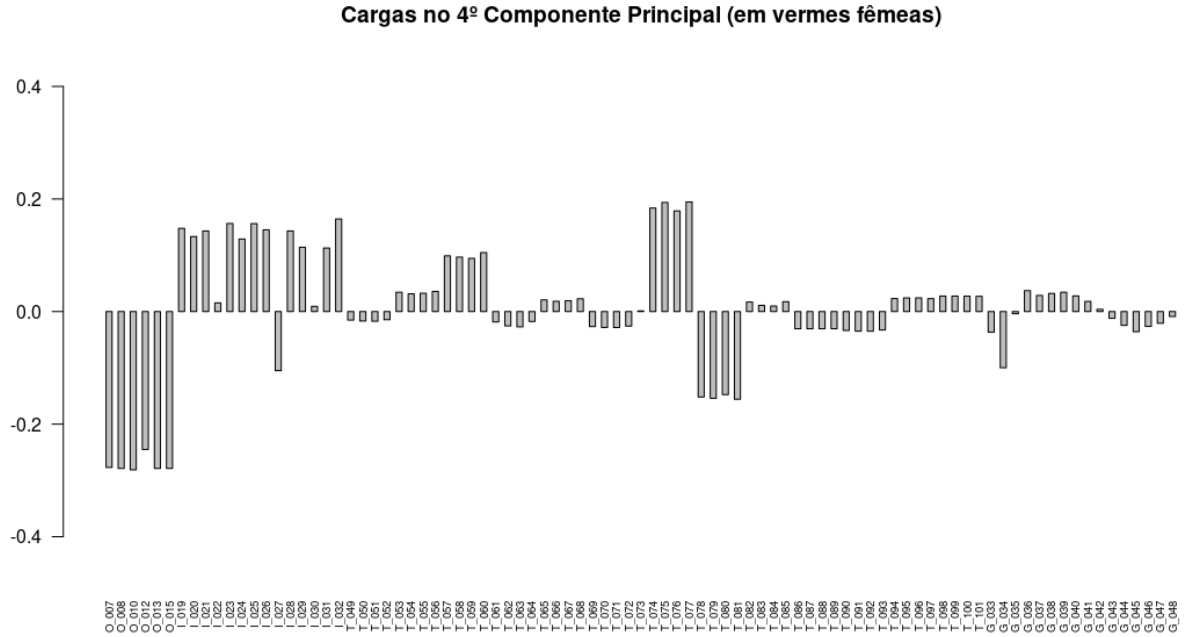

Variáveis Originais
T_049
T_050
T_051
T_052
T_053
T_054
T_055
T_056
T_057
T_058
T_059
T_060
T_061
T_062
T_063
T_064
T_065
T_066
T_067
T_068
T_069
T_070
T_071
T_072
T_073
T_074
T_075
T_076
T_077
T_078
T_079
T_080
T_081
T_082
T_083
T_084
T_085
T_086
T_087
T_088
T_089
T_090
T_091
T_092
T_093
T_094
T_095
T_096
T_097
T_098
T_099
T_100
T_101
frequência
Medida

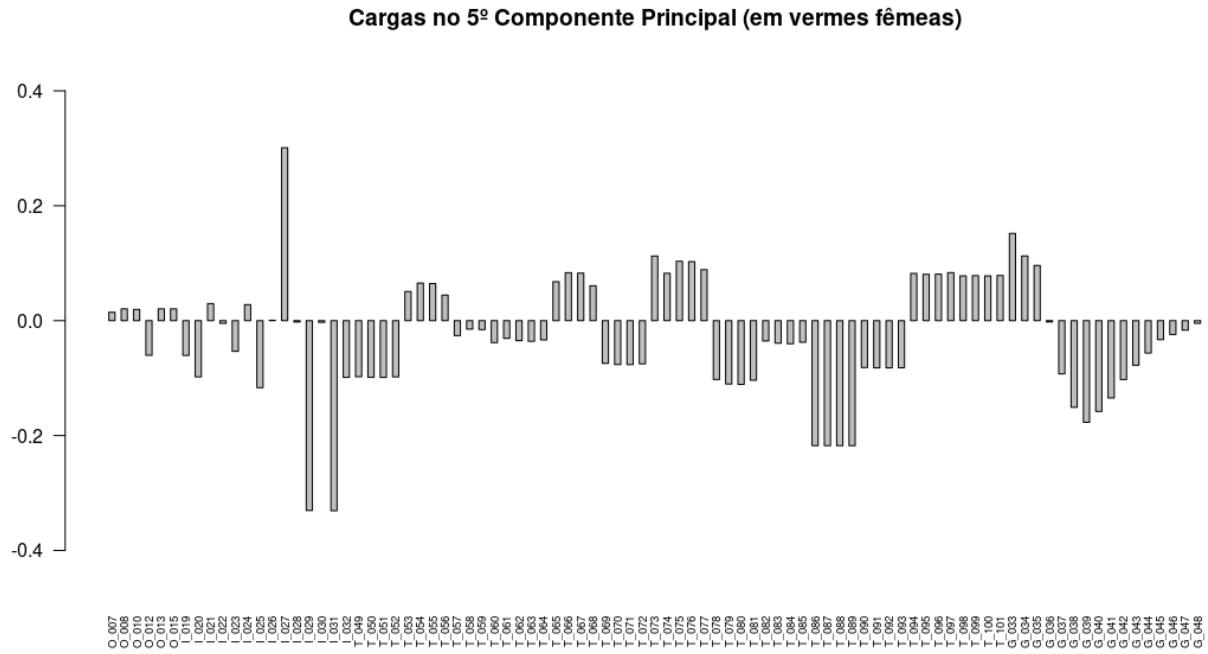

Variáveis Transformadas
T_049
T_050
T_051
T_052
T_053
T_054
T_055
T_056
T_057
T_058
T_059
T_060
T_061
T_062
T_063
T_064
T_065
T_066
T_067
T_068
T_069
T_070
T_071
T_072
T_073
T_074
T_075
T_076
T_077
T_078
T_079
T_080
T_081
T_082
T_083
T_084
T_085
T_086
T_087
T_088
T_089
T_090
T_091
T_092
T_093
T_094
T_095
T_096
T_097
T_098
T_099
T_100
T_101
frequência
Medida

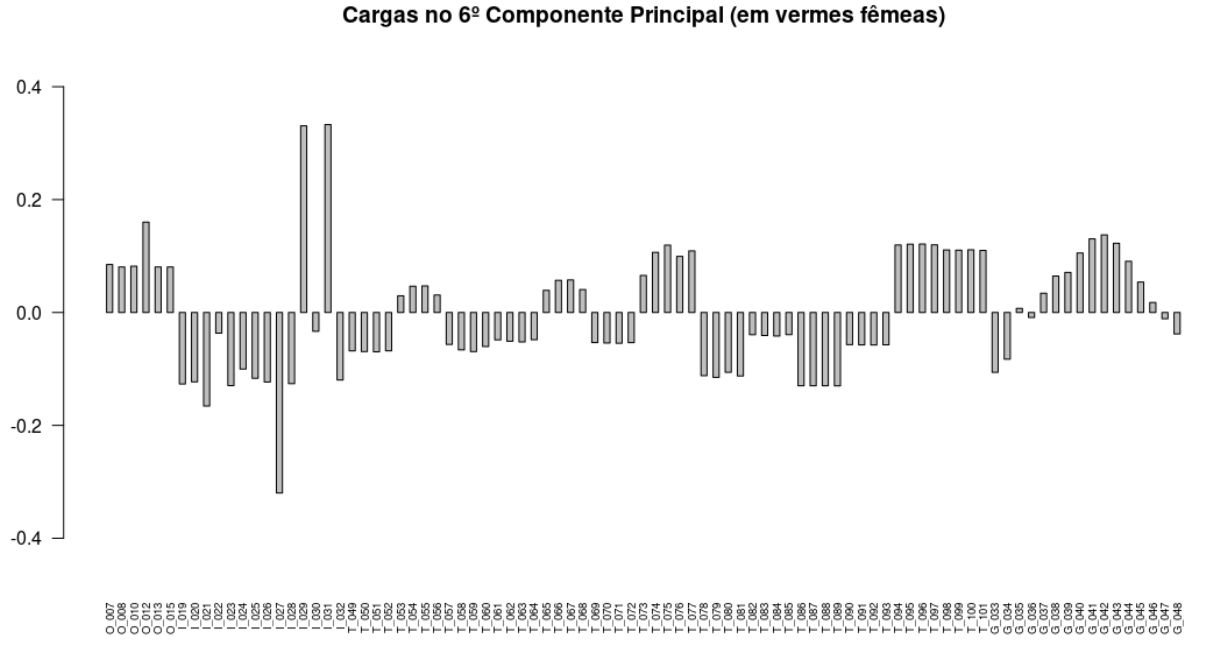

Variáveis Originais
Variáveis Transformadas
G_033
G_034
G_035
G_036
G_037
G_038
G_039
G_040
G_041
G_042
G_043
G_044
G_045
G_046
G_047
G_048
frequência
Medida

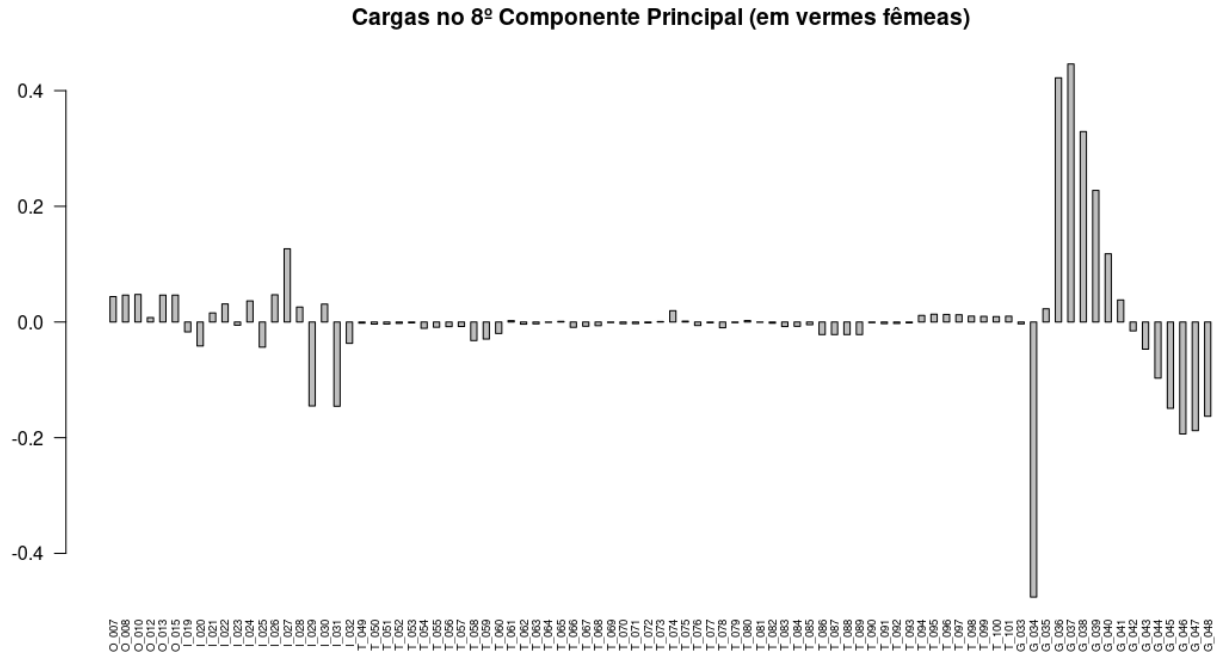
7.5 Anexo 5: Gráficos de barras das cargas nos primeiros componentes principais dos dados de pré-tratamento

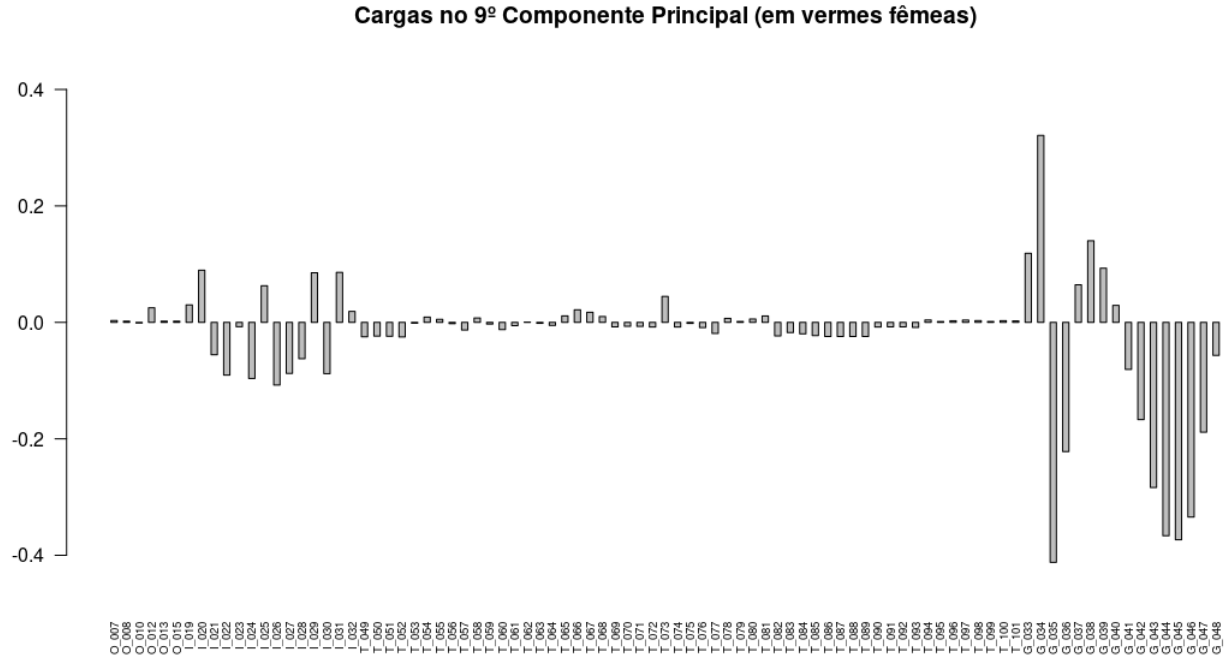




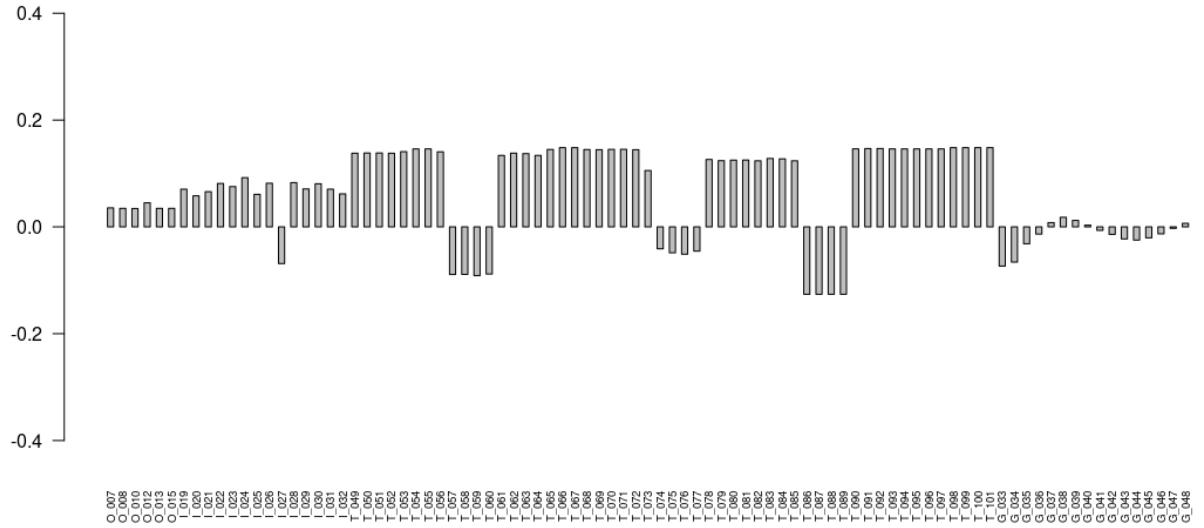

Cargas no 5º Componente Principal (em vermes fêmeas)
0.4
0.2
0.0
-0.2
-0.4
O_007
O_008
O_010
O_012
O_013
O_015
I_019
I_020
I_021
I_022
I_023
I_024
I_025
I_026
I_027
I_028
I_029
I_030
I_031
I_032
T_049
T_050
T_051
T_052
T_053
T_054
T_055
T_056
T_057
T_058
T_059
T_060
T_061
T_062
T_063
T_064
T_065
T_066
T_067
T_068
T_069
T_070
T_071
T_072
T_073
T_074
T_075
T_076
T_077
T_078
T_079
T_080
T_081
T_082
T_083
T_084
T_085
T_086
T_087
T_088
T_089
T_090
T_091
T_092
T_093
T_094
T_095
T_096
T_097
T_098
T_099
T_100
T_101
G_033
G_034
G_035
G_036
G_037
G_038
G_039
G_040
G_041
G_042
G_043
G_044
G_045
G_046
G_047
G_048


Cargas no 6º Componente Principal (em vermes fêmeas)
0.4
0.2
0.0
-0.2
-0.4
O_007
O_008
O_010
O_012
O_013
O_015
I_019
I_020
I_021
I_022
I_023
I_024
I_025
I_026
I_027
I_028
I_029
I_030
I_031
I_032
T_049
T_050
T_051
T_052
T_053
T_054
T_055
T_056
T_057
T_058
T_059
T_060
T_061
T_062
T_063
T_064
T_065
T_066
T_067
T_068
T_069
T_070
T_071
T_072
T_073
T_074
T_075
T_076
T_077
T_078
T_079
T_080
T_081
T_082
T_083
T_084
T_085
T_086
T_087
T_088
T_089
T_090
T_091
T_092
T_093
T_094
T_095
T_096
T_097
T_098
T_099
T_100
T_101
G_033
G_034
G_035
G_036
G_037
G_038
G_039
G_040
G_041
G_042
G_043
G_044
G_045
G_046
G_047
G_048

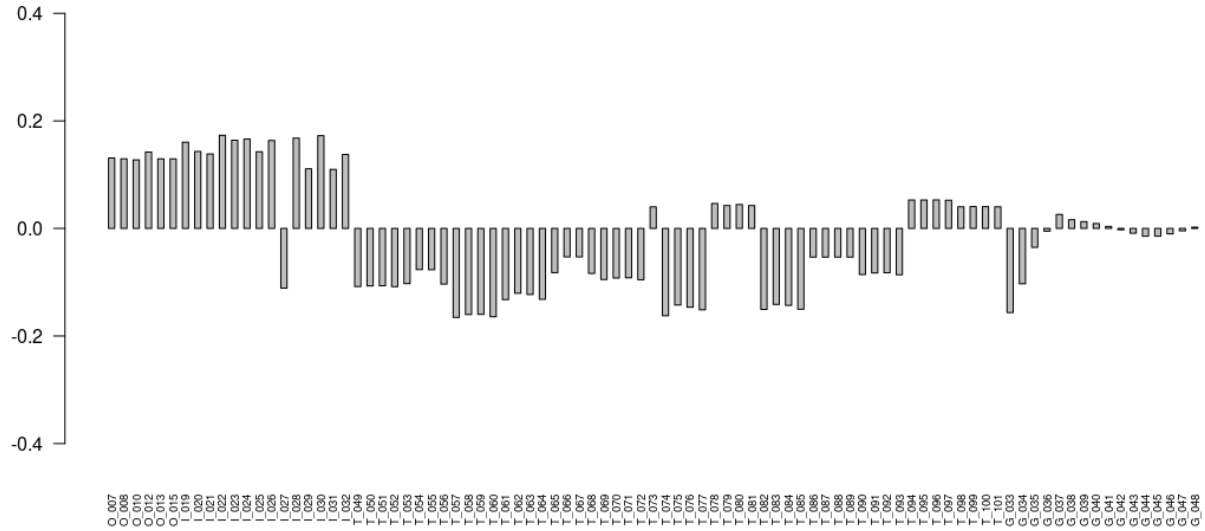

Cargas no 8º Componente Principal (em vermes fêmeas)
0.4
0.2
0.0
-0.2
-0.4


Cargas no 9º Componente Principal (em vermes fêmeas)
0.4
0.2
0.0
-0.2
-0.4

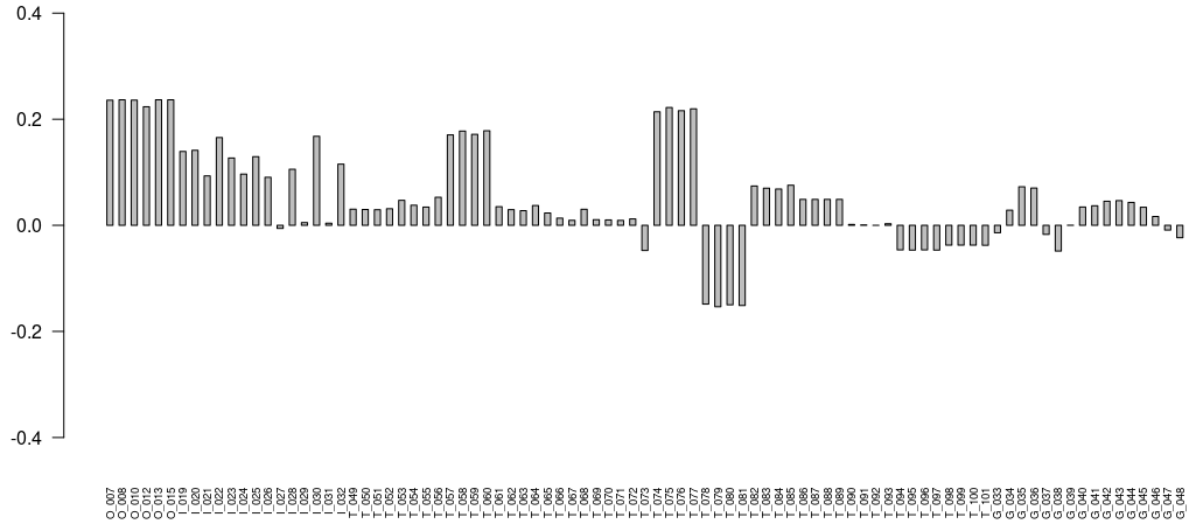
Cargas no 1º Componente Principal (em vermes machos)



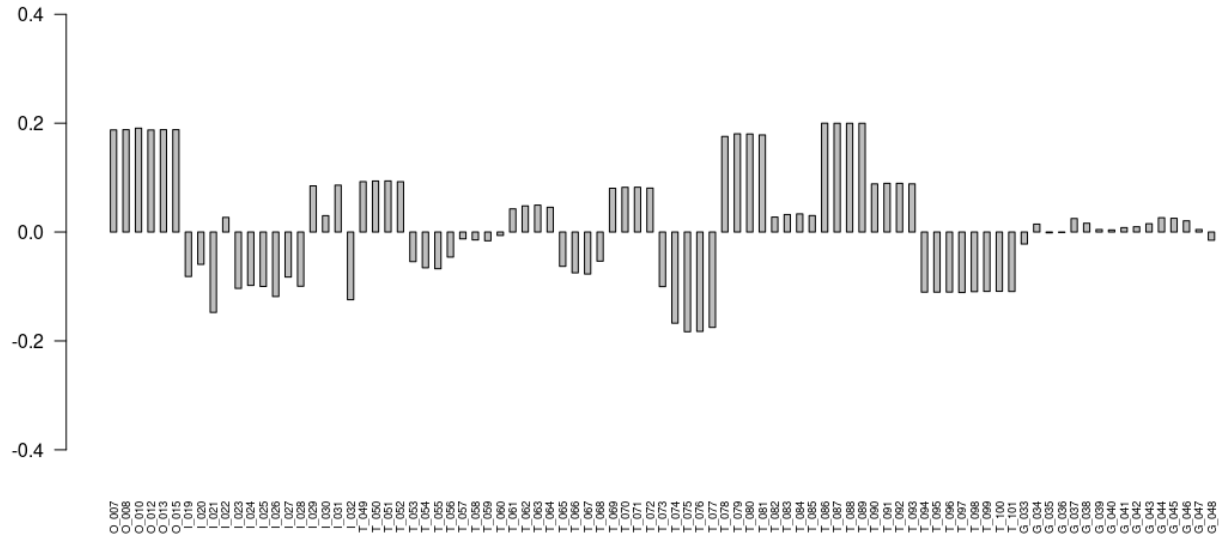
Cargas no 2º Componente Principal (em vermes machos)



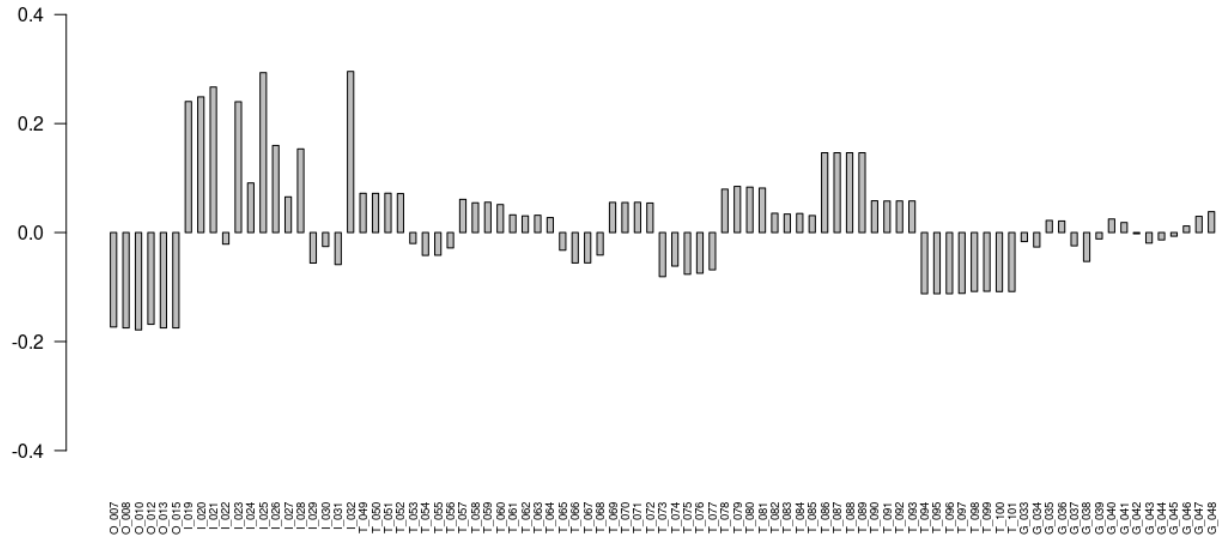
Cargas no 3º Componente Principal (em vermes machos)



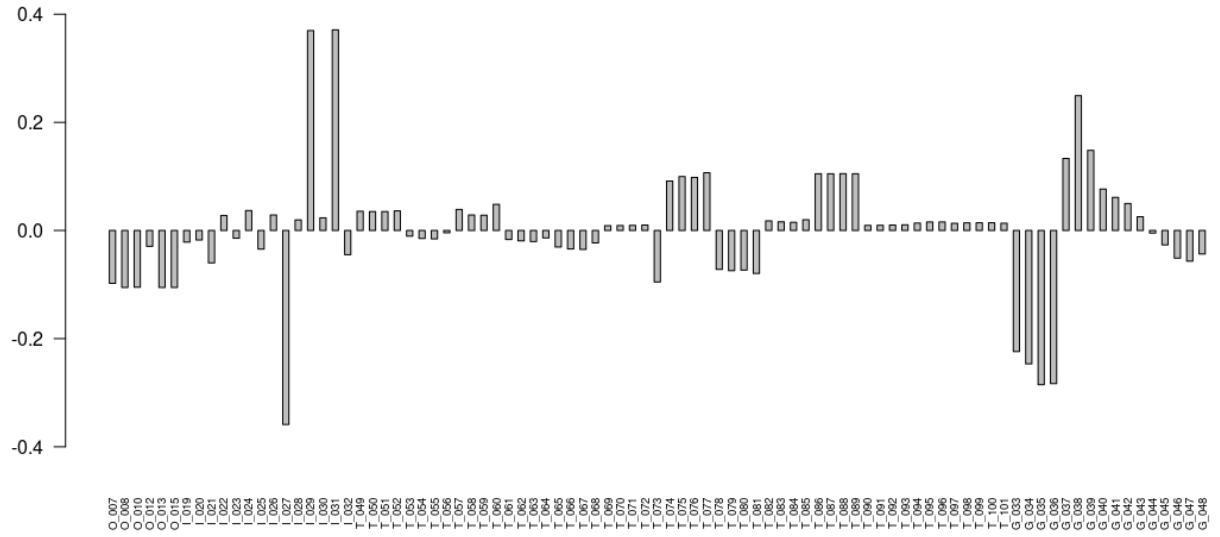
Cargas no 4º Componente Principal (em vermes machos)



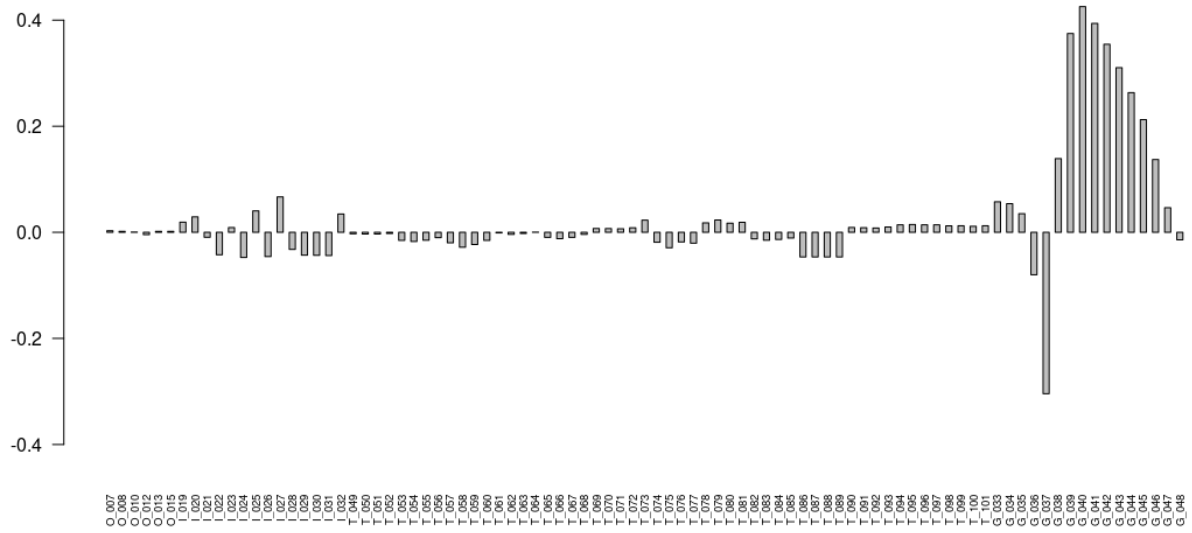
Cargas no 5º Componente Principal (em vermes machos)



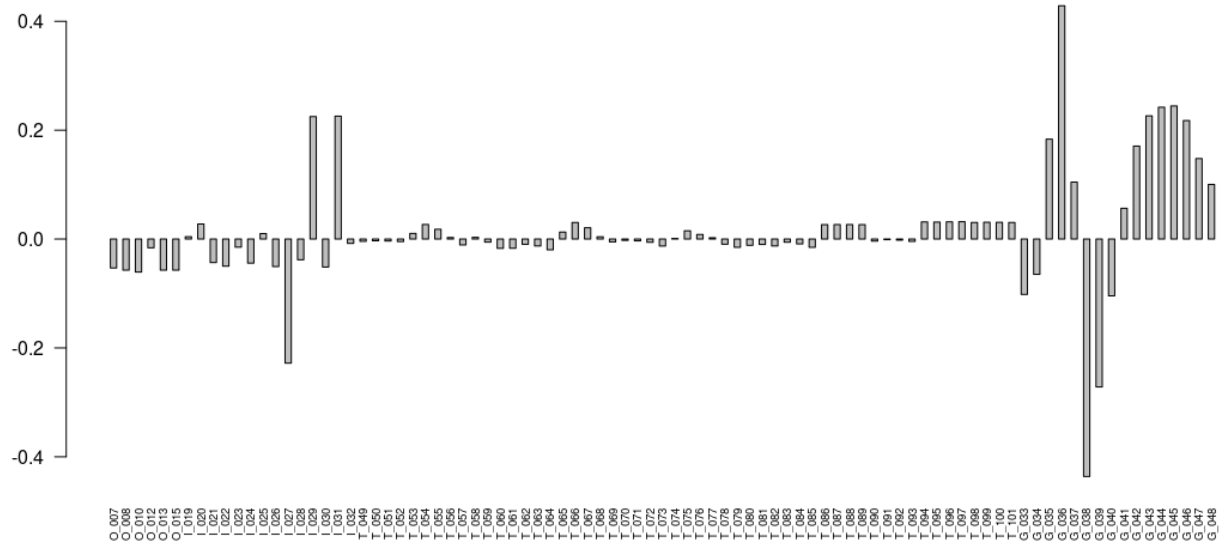
Cargas no 6º Componente Principal (em vermes machos)

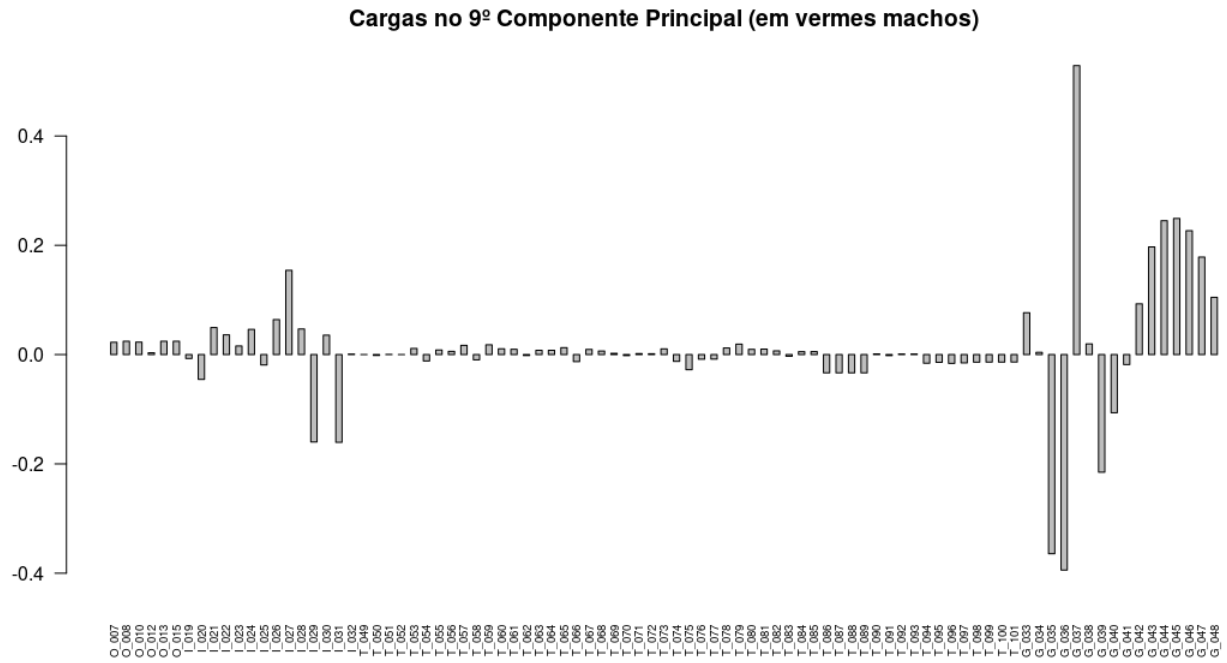


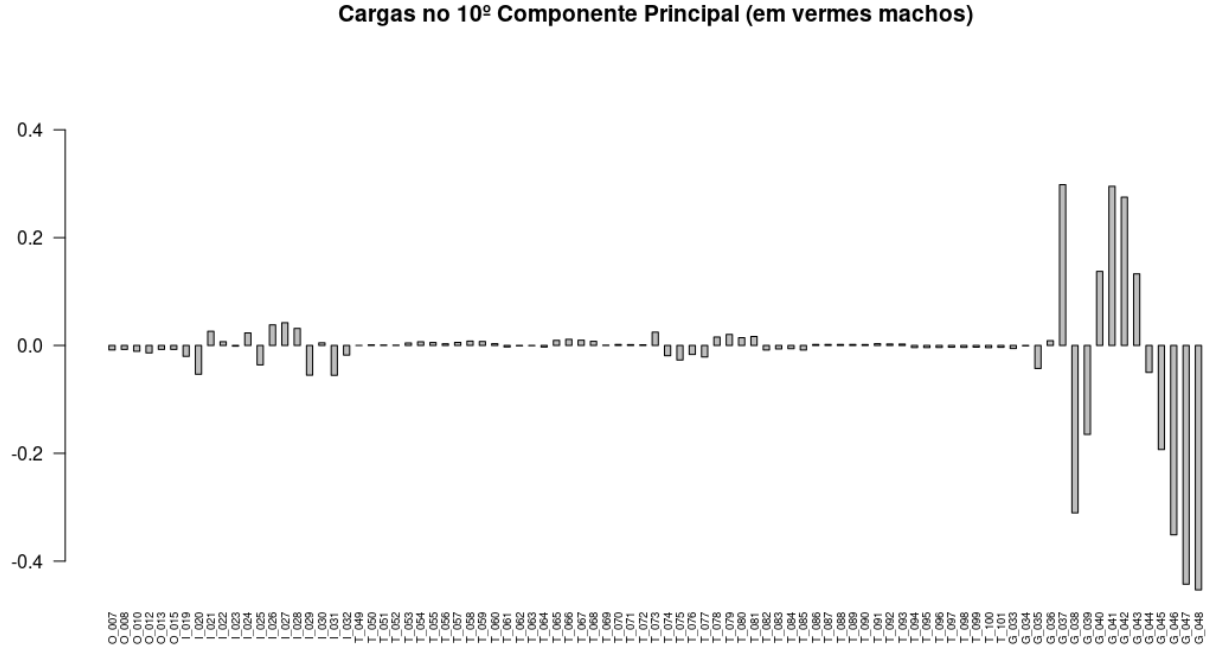
Cargas no 7º Componente Principal (em vermes machos)



Cargas no 8º Componente Principal (em vermes machos)




Cargas no 9º Componente Principal (em vermes machos)
0.4
0.2
0.0
-0.2
-0.4
O_007 O_008 O_010 O_012 O_013 O_015 I_019 I_020 I_021 I_022 I_023 I_024 I_025 I_026 I_027 I_028 I_029 I_030 I_031 I_032 T_049 T_050 T_051 T_052 T_053 T_054 T_055 T_056 T_057 T_058 T_059 T_060 T_061 T_062 T_063 T_064 T_065 T_066 T_067 T_068 T_069 T_070 T_071 T_072 T_073 T_074 T_075 T_076 T_077 T_078 T_079 T_080 T_081 T_082 T_083 T_084 T_085 T_086 T_087 T_088 T_089 T_090 T_091 T_092 T_093 T_094 T_095 T_096 T_097 T_098 T_099 T_100 T_101 G_033 G_034 G_035 G_036 G_037 G_038 G_039 G_040 G_041 G_042 G_043 G_044 G_045 G_046 G_047 G_048


Cargas no 10º Componente Principal (em vermes machos)
0.4
0.2
0.0
-0.2
-0.4
O_007 O_008 O_010 O_012 O_013 O_015 I_019 I_020 I_021 I_022 I_023 I_024 I_025 I_026 I_027 I_028 I_029 I_030 I_031 I_032 T_049 T_050 T_051 T_052 T_053 T_054 T_055 T_056 T_057 T_058 T_059 T_060 T_061 T_062 T_063 T_064 T_065 T_066 T_067 T_068 T_069 T_070 T_071 T_072 T_073 T_074 T_075 T_076 T_077 T_078 T_079 T_080 T_081 T_082 T_083 T_084 T_085 T_086 T_087 T_088 T_089 T_090 T_091 T_092 T_093 T_094 T_095 T_096 T_097 T_098 T_099 T_100 T_101 G_033 G_034 G_035 G_036 G_037 G_038 G_039 G_040 G_041 G_042 G_043 G_044 G_045 G_046 G_047 G_048