Visual-its-Better-HWE:

Pipeline básico para visualização de dados obtidos no PopGen Fishpond

Autor: Samuel Chagas de Assis

Resumo: A visualização de dados é uma etapa importante para interação e relação de dados obtidos em experimentos. Na genética populacional, a interpretação dos dados de proporção e sua aplicação são um grande desafio no processo de aprendizagem dos estudantes de graduação. Por isso, ferramentas de interface amigável são desenvolvidas para facilitar a interação, cruzamento e visualização de dados genéticos de maneira didática. Nesse artigo, apresenta-se uma *pipeline*, baseado em linguagem de programação R, para visualização de dados obtidos nos experimentos da interface web *PopGen Fishpond,* buscando otimizar o intercruzamento das variáveis geradas pelo expertimento virtual e aprimorar o processo de aprendizagem. Por fim, a visualização é incrementada com um teste clássico para o equilíbrio de Hardy-Weinberg , onde o *Qui-quadrado* e *p-valor* auxiliam na comparação entre diferentes tratamentos.

Palavras-chave: *HardyWeinberg, Bioinformática, Evolução*.

1. Introdução

O avanço de tecnolgias de genotipagem e sequenciamento alimentam continuamente o armazenamento de dados para pesquisa de genética populacional, possibilitando a expansão de variáveis coletadas em um tempo reduzido, como é o caso dos Estudos de Associação de Genômica Ampla (ou *GWAS*, sigla em inglês) (BOUGHTON et al., 2020). No entanto, o teste das proporções do equlíbrio de Hardy-Weingberg (HWE), apesar da sua simplicidade em relação a abordagens recentes, continua sendo fundamental nos avanços da área. Sua aplicação é destacada nas primeiras análises de triagem amostral atualmente e sua importância é refletida nas inúmeras ferramentas que aplicam os príncipios de HWE, especialmente em pacotes de linguagem em Python e R, como Genopop para Python e *genetics*, *adegenet*, *GAP* e *HardyWeinberg* para R (SANTOS et al., 2020).

No entanto, essa abordagem não é limitada ao campo da pesquisa científica, a licenciatura incorpora os príncipios de HWE no ensino da Evolução, utilizando a genética populacional para contextualizar genética e evolução, como: a dinâmica de genes dentro de uma população diferem dos genes herdados; foco na frequência alélica, em oposição aos indíviduos; e exemplos relevantes para demonstração do HWE (BREWER et al., 2013). Por isso, esse trabalho apresenta uma ferramenta para facilitar a visualização de dados obtidos através do *PopGen Fishpond* (SUSILAWATI et al., 2019), utilizando o pacote *R HardyWeinberg* e testando através de diferentes tratamentos ao longo das gerações de carpas Koi.

1.1 Equilíbrio de Hardy-Weinberg

Os princípios de Hardy-Weinberg define que frequências alélicas e fenotípicas permanecem constante ao longo das gerações com ausência de processo evolutivo, sua importância está em avaliar se existem influênias evolutivas atuando sob determinado marcador. A partir do ponto de vista matemático, descreve-se que pra determinados marcadores dialélicos, como A e B, suas respectivas frequências p e q $(p + q = 1)$ estão em HWE se as frequências relativas fenotípicas, como f_{AA}, f_{AB} e f_{BB} são dadas por p^2, $2pq$ e q^2, respectivamente. Além disso, essa abordagem pode ser seguida para sistemas com multiplos alelos A_i,... A_k com frequências p_1,...., $p_k$considerando a frequência para homozigotos $p_i^{\,2}$ e heterozigotos $2p_ip_j$ (GRAFFELMAN, 2015).

O HWE pressupões que uma população esteja em equilibrio quando: (I) não há mutação, (II) não migração,ou seja, sem fluxo gênico; (III) acasalamento aleatório; (IV) população significativamente grande ($N \to \infty$); (V) ausência de processo seletivo, todos os alelos possuem mesmo potencial de adaptação (SANTOS et al., 2020). Existem diversos testes estátisticos que avaliam se determinado marcador (i.e. genes HLA, ABO, etc) pode estar em HWE ou não. Uma abordagem clássica é o teste de Qui-quadrado (Equação 1) o qual compara a contagem de genotípos observadas (f_{AA}, f_{AB}, f_{BB}) com a contagem esperada em HWE ($e_{AA} = np^2$, $e_{AB} = 2np$, $e_{BB} = nq^2$, $n = número\ total\ de\ indíviduos$).

$$x^2 = \frac{(f_{AA} - e_{AA})^2}{e_{AA}} + \frac{(f_{AB} - e_{AB})^2}{e_{AB}} + \frac{(f_{BB} - e_{BB})^2}{e_{BB}} \quad \text{(Eq. 1)}$$

Além disso, demais aboragens estatísticas são aplicadas para evitar distúrbios durante a coleta de dados e análise, como as correções consiederadas na equação do Qui-quadrado (i.e. Correção de Yates, que avalia a independência de tabales de contigência) ou Teste exato para HWE de significância estatística.

1.3 Pacote *HardyWeinberg*

O Pacote HardyWeinberg foi desenvolvido em 2015 e consiste em uma ferramenta para análise de marcadores genéticos dialélicos, ambientada por linguagem R e utilizada na representação gráfico do (des)equiĺibrio de Hardy-Weinberg, como gráfico ternário e Q-Q (GRAFFELMAN, 2015). O pacote está disponibilizado em https://cran.r-project.org/web/packages/HardyWeinberg/index.html.

2. Metodologia

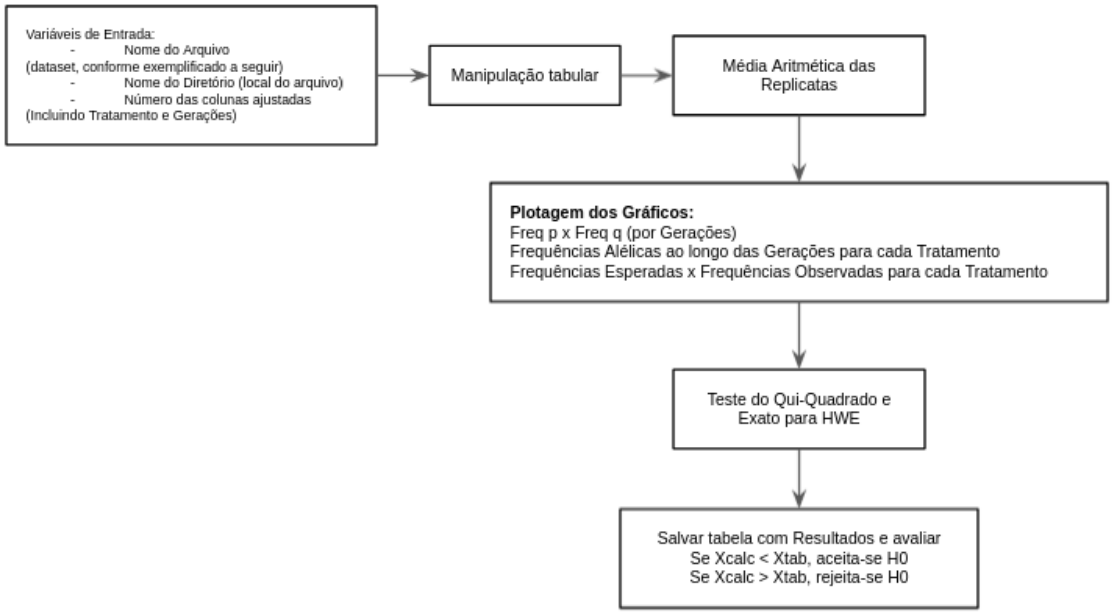
Os métodos compreendem a construção do *Pipeline* em linguagem R e avaliação através de diferentes tratamentos.

2.1 Construção do Visual-its-Better-HWE

O processo de construção partiu de um *dataset* composto por dados adquiridos no *PopGen Fishpond*, compostos pelos parâmetros utilizados no design do experimento e os resultados ao longo das gerações, conforme o Quadro 1 no Material Suplementar (MS). A visualização dos resultados foi obtido através de um *pipeline* basedo em R (R versão 4.1.1 e RStudio 1.4). O código e o *dataset* utilizado e resultados dos experimentos estão disponíveis em https://github.com/chagas98/VisualItsBetterHWE (Dados.csv para variáveis de entrada e resultados dos experimentos; Evolucao.R para o código; Resultados_cn.csv e Resultados_fit.csv para resultados do Qui-Quadrado). Logo

após, aplicou-se a avaliação do funcionadmento do programa. O pipeline é organizado conforme demonstrado no Fluxograma 1.

Fluxograma 1 - Visual-its-Better-HWE Pipeline



2.2 Avaliação do Pipeline

2.2.1 Grupo Controle

Inicialmente, foi gerado um grupo controle no Fishpond, ou seja, um grupo que se encontra em condições baseadas no HWE. Por isso, devido a falta de informação do software, avaliou-se qual a taxa de mortalidade e tamanho da prole ideal para se ter as condições de HWE bem ajustadas e a geração de um controle adequado para as futuras análises. A coleta dos dados ocorreu ao longo de 100 gerações com intervalo de 10 gerações. Os tratamentos ocorreram em triplicata. Nas configurações do *Fishpond*, foram escolhidos valores que excluíriam migração, seleção, mutação e avalia-se aqui a mortalidade da população, os valores podem ser conferidos a Tabela 1 no MS.

2.2.2 Seleção vs HWE

Como forma de avaliar a efitividade dos gráficos gerados, buscou-se verificar se a mudança da taxa adaptativa, conferida para cada genótipo, poderia ser avaliada por diferentes abordagens gráficas, facilitando a inferência no relacionamento dos dados e na constatação do desequlíbrio. Sendo assim, foi configurado um *fitness* igual a 0.5 para os genotipos rr, RR, Rr, separadamente, e monitorados ao longo de 100 gerações com intervalo de 20 gerações. O restante das configurações foram ajustadas de acordo com os grupos avaliados anteriormente, Cn e Cnn (FitrrCn, FitRRCn, FitRrCn e FitrrCnn, FitRRCnn, FitRrCnn). Logo após, os gráficos gerados foram comparados com os resultados estatísticos. Todo o experimento ocorreu em triplicata.

3. Resultados

3.1 Grupo Controle

A definição do grupo controle foi avaliada comparado dois tratamento: a taxa de mortalidade igual o tamanho da prole dos peixes (Cn) e o tamanho da prole três vezes maior que a taxa de mortalidade (Cnn). Observou-se um distanciamento da proporção alélica inicial no grupo Cn para os alelos r (q) e R (p) , podendo indicar um processo adaptativo (Figura 1).

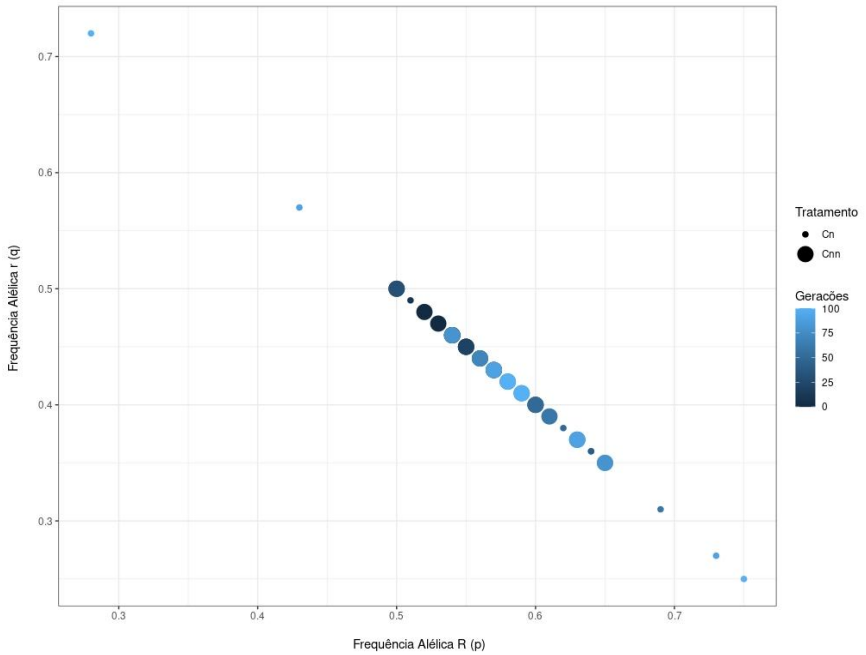


Figura 1 - Relação Frequência q (alelo r) x Frequência p (alelo R.

Curiosamente, a frequência alélica ao longo do tempo se mostrou mais estável graficamente para o grupo Cn comparado ao grupo Cnn. Apesar de ter gerado um ajuste baseado em regressão linear entre as replicatas, percebe-se que há uma discrepância entre os dados obtidos a partir do grupo Cn (Figura 2).

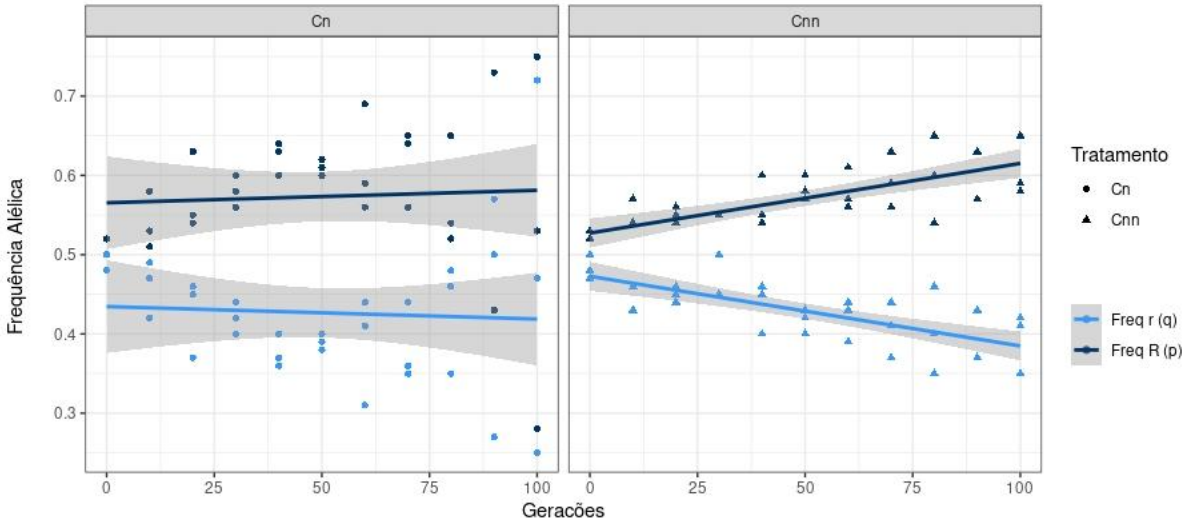


Figura 2 - Frequência alélica ao longo do tempo. Linhas: replicatas ajustadas com modelo de regressão linear

Além disso, quando comparado ao gráfico de HWE (onde se encontram valores esperados para frequência genotípica em HWE), observa-se que o grupo Cn

possui maior discrepância (Figura 3). Portanto, para melhor avaliar essa relação, a análise de Qui-quadrado foi realizada posteriormente.

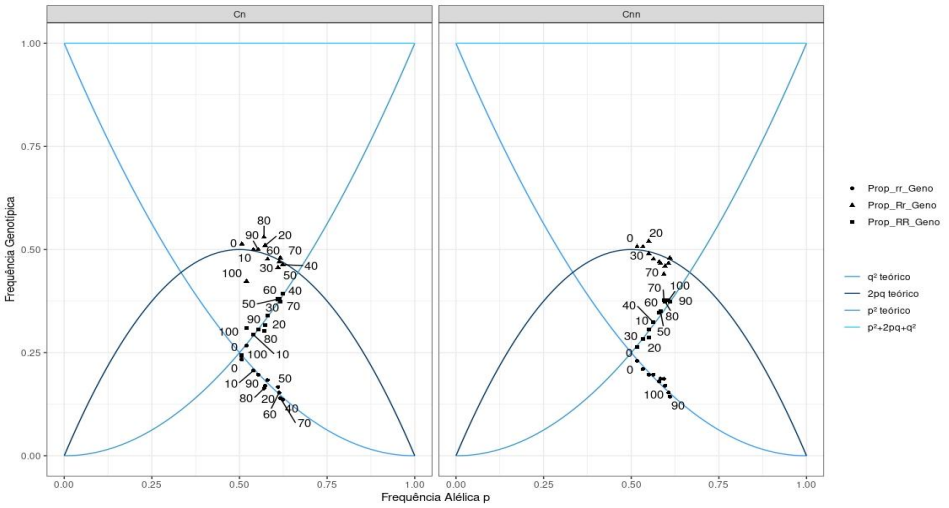


Figura 3 - Relação Frequência q x Frequência p. Prop_rr_Geno, Prop_Rr_Geno Prop_RR_Geno: Proporção do genotipo rr, Rr, RR, respectivamente.

Por meio dos cálculos descritos no item 1.1, obteve-se os valores de Qui-quadrado para cada uma das gerações. Na tabela 2 (no MS), observa-se que Cn e Cnn tiveram resultados semelhantes em relação a aceitação ou rejeição da hipótese nula. Na aba Resultados, valores Xcalc superiores ao Xtab (~3.841, em um grau de liberdade = 1) foram Rejeitados e os inferiores foram aceitos (H0). Portanto, utilizou-se ambos os grupos como controle para as próximas análises.

3.2 Seleção vs HWE

Conforme constatado anteriormente, a alteração da taxa de mortalidade em relação ao tamanho da prole gerou um distanciamento das proporções alélicas iniciais. Como avaliado na Figura 4, mesmo com alteração de *fitness* dos genótipos, o grupo Cn contribui para um possíel processo adaptativo. Curiosamente, algumas frequências alélicas para genótipos *rr* com *fitness* reduzido apresentaram altas frequências.

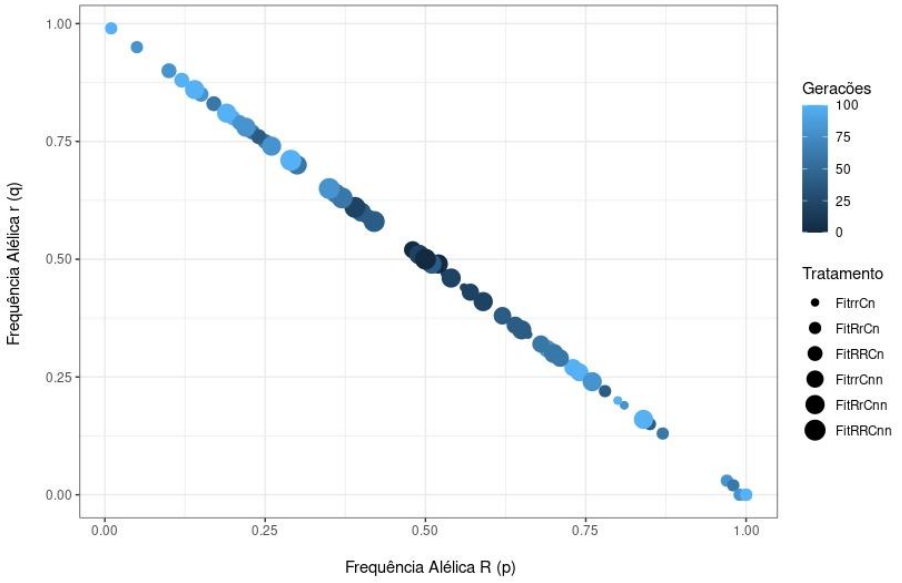


Figura 4 - Relação Frequência q x Frequência p. Tratamento: FitrrCn - *Fitness rr = 0.5 para Cn;* FitRrCn - *Fitness Rr = 0.5 para Cn;* FitRRCn - *Fitness RR = 0.5 para Cn;* FitrrCnn - *Fitness rr = 0.5 para Cnn;* FitRrCnn - *Fitness Rr = 0.5 para Cnn;* FitRRCn - *Fitness RR = 0.5 para Cnn.*

No entanto, a Figura 2 contribui para elucidar que as frequências elevadas de *rr, observadas na Figura 1,* possuem contribuição da alta variabilidade encontrada quando reduzido o fitness heterozigoto no grupo Cn (FitRrCn). Além disso, constata-se que o tratamento Cn (taxa de mortalidade = tamanho da prole) contribui conseideravelmente no processo adaptativo (Figura 5).

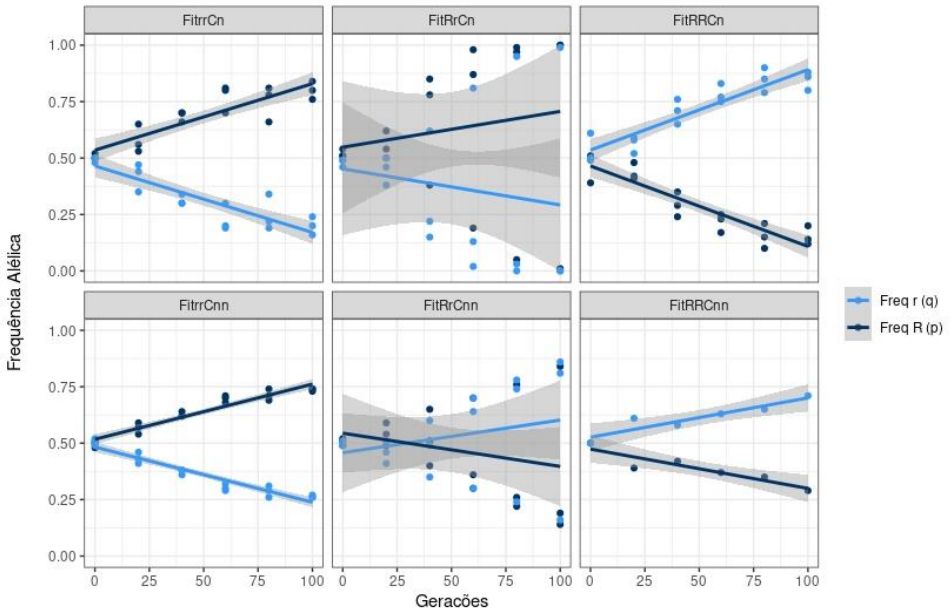


Figura 5 - Frequência alélica ao longo do tempo.

No entanto, ao analisar as diferentes frequências genotípicas em relação as frequências alélicas esperadas, ambos os tratamentos tornam dificíl a visualização de quais gerações estão em desequilíbrio (Figura 6).

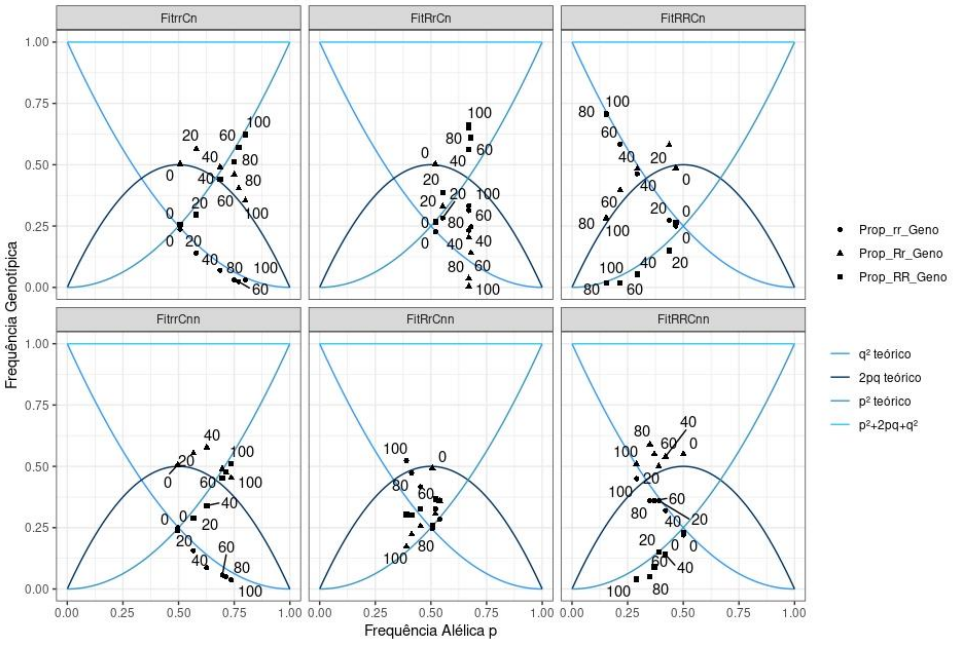


Figura 3 - Relação Frequência q x Frequência p. Prop_rr_Geno, Prop_Rr_Geno Prop_RR_Geno: Proporção do genótipo rr, Rr, RR, respectivamente.

Por fim, o teste de Qui-quadrado apresentou rejeição singnificativamente em em todos os tratamentos dessa seção. As gerações iniciais apresentaram equilíbrio inicial, evento esperado já que se configura a proporção alélica inicial.

Discussão

No experimento, os alelos R e r, que determinam a pigmentação das carpas Koi, foram avaliados em função da alteração da taxa de mortalidade da população em relação ao tamanho da prole e logo após, avaliou-se a influência da taxa adaptativa (fitness) conferida por cada genótipo. O pipeline desenvolvido conseguiu relacionar as variáveis em diferentes perspectivas, possibilitando uma percepção mais fundamentada sobre a inferência do HWE.

Mesmo que em taxas iguais, a população de peixes reduziu drasticamente, sugerindo que a taxa de mortalidade no Fishpond apresenta um influência superior

ao tamanho da prole durante a simulação. Essa relação não distanciou significativamente a população de carpas do HWE, o qual foi avaliado graficamente e por meio do teste Qui-quadrado com p valor inferior ao intervalo de confiança na maioria dos casos. Em contrate, a diferença entre essas duas váriaveis não incrementaram o HWE, sugerindo que a mortalidade exerce pode ter uma influência baixa sobre o processo de seleção. No entanto, maiores análises são necessárias para avaliar a taxa de mortalidade superior ao tamanho da prole.

Em relação ao *fitness* dos genótipos, observou-se que essas váriaveis influênciam consideravelmente no desequilibrio da população de carpas e quando somado as influências da taxa de mortalidade, podem acelerar o processo adaptativo.

Por fim, essa ferramenta requer maiores incrementos, especialmente do ponto de vista estatístico, podendo gerar análises mais robustas em relação ao Qui-Quadrado, inserindo determinadas correções e possibilitem maiores intercruzamentos de variáveis.

Referências

BOUGHTON, A. P. et al. Exploring and visualizing large-scale genetic associations by using PheWeb. Nature Genetics: Correspondence, 2020.

BREWER, M. S.; GARDNER, G. E. Teaching Evolution through the Hardy-Weinberg Principle: A real-time, active-learning exercise using classroom response devices. The American Biology Teacher, 2013.

GRAFFELMAN, J. Exploring Diallelic Genetic Markers: The HardyWeinberg Package. Journal of Statistical Software, 2015.

SANTOS, F. A. HW_TEST, a program for comprehensive HARDY-WEINBERG equilibrium testing. Evolutionary Genetics, 2020.

SUSILAWATI, P. R. Implementation of Web-Based Virtual Laboratory Media in Learning

Material Suplementar

Quadro 1 - Dicionário de Dados

Colunas	**Descriçao**
Name	Nome do Experimento
Tratamento	Nome do Tratamento
Initial_size	Tamanho da população inicial
Capacity	Capacidade populacional
SexRatio	Razão entre machos e fêmeas
Mortality	Mortalidade da População
Brood	Tamanho da prole
RpropInit	Proporção Alélica R inicial
Migration	Taxa migratória
MigRprop	Proporção Alélilca R de imigrantes
Rtor	Taxa de mutação de R para r

rtoR	Taxa de mutação de r para R
Fitnrr	Adaptação para rr
FitnRr	Adaptação para Rr
FitnRR	Adaptação para RR
Generation	Gerações
Population_Size	Tamanho da População
Prop_R_Al	Proporção Alélica R
Prop_r_Al	Proporção Alélica r
Prop_rr_Geno	Proporção Genotípica rr
Prop_Rr_Geno	Proporção Genotípica Rr
Prop_RR_Geno	Proporção Genotípica RR

Tabela 1 - Desenho Experimental

Condições	**Cn**	**Cnn**
Initial size	500	500
Carrying Capacity	500	500
Mortality Rate	10	5
Brood Size	10	15
Initial R Allele Prop	0.5	0.5
Migration Rate	0	0
MIgrant R allele	1	1
R to r	0	0
r to R	0	0
Fit rr	1	1
Fit Rr	1	1

Fit RR	1	1
Strength of Assortment	0	0

Tabela 2 - Resultados Qui-Quadrado (Cn/Cnn)

	Tratamento	Generation	RRN	RrN	rrN	Xcalc	pval	Resultado
1	Cn	0	122	257	117	0.552862690300262	0.4571503539112	H0
2	Cnn	0	132	254	115	0.0743648883347021	0.785084400866084	H0
3	Cn	10	114	194	80	0.00635288543693749	0.936471890606877	H0
4	Cnn	10	154	247	99	0.00485691627268765	0.944439143961186	H0
5	Cn	20	93	150	50	0.497248104875283	0.48071178305042	H0
6	Cnn	20	144	262	99	0.931074114577636	0.334583972539176	H0
7	Cn	30	85	119	46	0.0825430004745369	0.773880343549317	H0
8	Cnn	30	142	255	105	0.171860370813897	0.678463702250645	H0
9	Cn	40	89	105	31	0.0125051405697514	0.910961065864013	H0

10	Cnn	40	163	240	99	0.313578415845215	0.575492582968553	H0
11	Cn	50	86	103	38	0.421224954081303	0.516326448040278	H0
12	Cnn	50	175	234	94	0.876223206 29682	0.349238203555385	H0
13	Cn	60	95	117	38	0.010832264 7341548	0.917107317384801	H0
14	Cnn	60	174	235	90	0.378573345925553	0.538367748074122	H0
15	Cn	70	92	118	34	0.084644185 6803136	0.771099702955151	H0
16	Cnn	70	189	221	94	3.866966967232	0.049245438 7673331	Rejeita
17	Cn	80	77	135	42	1.526943339 48738	0.216571775280433	H0
18	Cnn	80	187	231	85	0.764400418504377	0.381954921599874	H0
19	Cn	90	71	116	45	0.009862801 11828439	0.920890839606438	H0
20	Cnn	90	186	239	71	0.115237806188977	0.734257994941833	H0
21	Cn	100	72	99	62	4.712055261 03763	0.029951822 7610657	Rejeita
22	Cnn	100	190	236	77	0.037419118 3026097	0.846614220051966	H0

**Rejeita: Xcalc > 3.841; H0: Xcalc< 3.841*

Tabela 3 - Resultados Qui-Quadrado (Fit)

Tratamento	Generation	Population_Size	Prop_R_Al	RRN	RrN	rrN	Xcalc	pval	Resultado
FitRRCn	0	493	0.4666666666666667	130	240	123	0.26424766495 6223	0.60721685511809 4	H0
FitRRCnn	0	503	0.5	116	277	111	4.67812156097 478	0.03054913585892 37	Rejeita
FitRrCn	0	500	0.52	133	252	113	0.05714304245 06147	0.81106982878586 4	H0
FitRrCnn	0	503.33333333333	0.5066666666666667	131	248	124	0.05854528473 94997	0.80881012795473 7	H0
FitrrCn	0	500	0.5066666666666667	128	252	118	0.04490700139 69102	0.83217512591488	H0
FitrrCnn	0	504.66666666666667	0.4966666666666667	121	256	126	0.11383412738 7614	0.73582055924194 6	H0
FitRRCn	20	283.66666666666667	0.4366666666666667	43	165	78	7.88553790043 032	0.00498316551881 038	Rejeita
FitRRCnn	20	503	0.39	75	252	181	0.59322605532	0.44117432758520	H0

							7402	4	
FitRrCn	20	213.33333333333	0.553333333333333	82	70	60	22.428923435502	2.18065281684281e-06	Rejeita
FitRrCnn	20	499.66666666666667	0.54	180	180	142	38.1645859843555	6.50218681911723e-10	Rejeita
FitrrCn	20	242.33333333333	0.58	72	137	34	5.42914087354972	0.0198034193640505	Rejeita
FitrrCnn	20	503	0.566666666666667	146	278	79	7.5215773939261	0.00609642729376754	Rejeita
FitRRCn	40	193.66666666666667	0.293333333333333	10	94	90	4.83152981771331	0.0279438247487228	Rejeita
FitRRCnn	40	500	0.42	70	270	160	6.38839730116459	0.0114868695783392	Rejeita
FitRrCn	40	163	0.67	92	33	38	46.5622529341272	8.87535442710156e-12	Rejeita
FitRrCnn	40	503	0.52	184	154	164	73.4402411807596	1.03729586555121e-17	Rejeita
FitrrCn	40	230.66666666666667	0.686666666666667	101	113	16	3.89700260251192	0.0483723373679508	Rejeita
FitrrCnn	40	502.33333333333333	0.626666666666667	171	290	44	25.0729218316446	5.52026219522115e-07	Rejeita
FitRRCn	60	169.66666666666667	0.216666666666667	3	67	99	4.12225440958247	0.0423226440027358	Rejeita
FitRRCnn	60	497	0.37	45	273	179	16.3510824943049	5.26260951077622e-05	Rejeita
FitRrCn	60	127.66666666666667	0.68	78	18	31	54.8623423462479	1.29274379889936e-13	Rejeita
FitRrCnn	60	502	0.453333333333333	164	129	209	115.118414792642	7.4135724612535e-27	Rejeita
FitrrCn	60	195.66666666666667	0.77	112	79	5	3.64792864368828	0.0561389892992633	H0
FitrrCnn	60	502.33333333333333	0.696666666666667	228	246	28	13.0546267622285	0.000302536117822392	Rejeita
FitRRCn	80	227.66666666666667	0.153333333333333	4	65	160	0.46938016354851	0.493272460556943	H0
FitRRCnn	80	496	0.35	25	293	179	44.9870395363994	1.98342838679962e-11	Rejeita
FitRrCn	80	110	0.67	72	4	34	89.3638903543183	3.28477744900494e-21	Rejeita
FitRrCnn	80	502.33333333333333	0.413333333333333	151	112	238	143.86830968365	3.79649672398 53e-33	Rejeita
FitrrCn	80	160.33	0.75	82	74	5	5.15392431482	0.02319389797893	Rejeita

		3333333333					397	22	
FitrrCnn	80	502.33333333333	0.713333333333333	239	238	25	12.048553615062	0.000518325814353088	Rejeita
FitRRCn	100	227	0.153333333333333	4	64	161	0.376715883125871	0.539366101892107	H0
FitRRCnn	100	507	0.29	20	259	228	25.599534752925 4	4.20140685335836e-07	Rejeita
FitRrCn	100	217.33333333333	0.67	144	1	72	209.085312385711	2.17483715160051e-47	Rejeita
FitRrCnn	100	501.33333333333	0.39	152	87	262	200.053640757562	2.03294809995623e-45	Rejeita
FitrrCn	100	177.33333333333	0.8	111	63	5	0.871311591406708	0.350592460472971	H0
FitrrCnn	100	503	0.736666666666667	257	228	18	13.990749352655 6	0.000183712273411545	Rejeita

**Rejeita: Xcalc > 3.841; H0: Xcalc< 3.841*