**Genotipagem de poliplóides:
um modelo de urnas e bolas**

Silvio Rodrigues de Faria Junior

TESE APRESENTADA
AO
INSTITUTO DE MATEMÁTICA E ESTATÍSTICA
DA
UNIVERSIDADE DE SÃO PAULO
PARA
OBTENÇÃO DO TÍTULO
DE
DOUTOR EM CIÊNCIAS

Programa: Estatística
Orientador: Profa. Dra. Júlia Maria Pavan Soler

Durante parte do período de desenvolvimento deste trabalho
o autor recebeu auxílio financeiro da CAPES e do CNPq

São Paulo, agosto de 2012

Genotipagem de poliplóides um modelo de urnas e bolas

Esta tese contém as correções e alterações
sugeridas pela Comissão Julgadora durante a defesa
realizada por Silvio Rodrigues de Faria Junior em 30/05/2012.

O original encontra-se disponível no Instituto de
Matemática e Estatística da Universidade de São Paulo.

Comissão Julgadora:

- Profa. Dra. Júlia Maria Pavan Soler (orientadora) - IME-USP
- Prof. Dr. Carlos Alberto de Bragança Pereira - IME-USP
- Prof. Dr. Antonio Augusto Franco Garcia - ESALQ-USP
- Prof. Dr. Luiz Lebensztajn - POLI-USP
- Prof. Dr. José Miguel Ortega - UFMG

Agradecimentos

Agradeço a Deus pela vida, aos meus pais pelo apoio incondicional, às minhas irmãs pelo incentivo, à minha namorada Rúbia pela compreensão, à Profa. Júlia Pavan Soler pela confiança no desenvolvimento deste trabalho, ao Prof. Antonio Augusto Garcia da Esalq e seu aluno Marcelo Molinari pelas orientações nos conceitos biológicos e pelo fornecimento de dados experimentais reais. À Capes e ao CNPq pelo apoio financeiro durante 18 meses do meu período de doutorado. Agradeço a USP e ao IME pela oportunidade de desenvolver meus estudos desde a graduação. E a incontáveis amigos, companheiros de jornada, e professores que possibilitaram meu desenvolvimento acadêmico.

Resumo

FARIA JR, S. R. **Genotipagem de poliplóides: um modelo de urnas e bolas**. 2012. 262 f. Tese (Doutorado) - Instituto de Matemática e Estatística, Universidade de São Paulo, São Paulo, 2012.

Desde os primórdios da agricultura e pecuária, o homem seleciona indivíduos com características desejáveis para reprodução e aumento da proporção de novos indivíduos com tais qualidades. Com o conhecimento da estrutura de DNA e o advento da engenharia genética, a identificação e caracterização de espécies e indivíduos conta com novas tecnologias para auxiliar no desenvolvimento de novas variedades de plantas e animais para diversos fins. Tais tecnologias envolvem procedimentos bioquímicos e físicos cada vez mais apurados que produzem medidas cada vez mais precisas, um exemplo disso são as técnicas que empregam a espectometria de massa para comparar polimorfismos de base única (SNPs). Nas plantas é comum a ocorrência de poliploidia, que consiste na presença de mais de dois cromossomos num mesmo grupo de homologia. A determinação do nível de ploidia é fundamental para a correta genotipagem e por consequência maior eficiência no estudo e aprimoramento genético de plantas. Neste trabalho fazemos uma caracterização do fenômeno da poliploidia com modelos probabilísticos de urnas e bolas, propondo um método eficiente e adequado de simulação, assim como uma técnica simples para inferir níveis de ploidia e classificar amostras bialélicas aproveitando características geométricas do problema. Análises de dados simulados e reais provenientes de um experimento de cana-de-açúcar foram realizadas com diferentes medidas de separação entre agrupamentos e diferentes condições experimentais. Para os dados reais, métodos gráficos descritivos evidenciam a corretude e coerência do método proposto, que pode ser generalizado para a genotipagem de locos multialélicos poliplóides. Todo código desenvolvido em linguagem R estão disponibilizados com o texto.

Palavras-chave: poliploidia, genotipagem de SNPs, modelos de urnas e bolas, simulação, agrupamentos.

Abstract

FARIA JR, S. R. **Polyploid genotyping: an urn model**. 2012. 262 f. Tese (Doutorado) - Instituto de Matemática e Estatística, Universidade de São Paulo, São Paulo, 2012.

Since the beginnings of agriculture and livestock, the man selects individuals with desirable characteristics to breed and increase the proportion of new individuals with such qualities. With knowledge of the DNA structure and the advent of genetic engineering, the identification and characterization of individual species can make use of new technologies to help develop new varieties of plants and animals for many purposes. These technologies involve complex biochemical and physical procedures that produce even more accurate measures, like techniques that employ mass spectrometry to compare single nucleotide polymorphisms (SNPs). In plants it is common the occurrence of polyploidy, which is the presence of more than two chromosomes in the same group of homology. The determination of polyploidy level is essential for correct SNPs genotype calling and therefore greater efficiency in the study and genetic improvement of plants. In this work we characterize the phenomenon of poliploidy with probabilistic urns and balls models, proposing an efficient and appropriate method of simulation, as well as a simple technique to infer ploidy levels and classify biallelic samples accurately taking advantage of geometrical characteristics of the problem. Analysis of simulated and real data from an experiment of sugarcane were conducted with different measures of separation between groups and different experimental conditions. For the actual data, descriptive graphical methods show the correctness and consistency of the proposed method, which can be generalized to multi-allelic loci genotyping polyploid. All code developed in language R are provided with the text.

Keywords: poliploidy, SNPs genotype calling, urns and balls models, simulation, clustering.

Sumário

1	Introdução	1
1.1	Considerações preliminares	1
1.2	Desafios e Objetivos	3
1.3	Organização do Trabalho	3
2	Poliploidia	4
2.1	Definição e origens	4
2.2	Poliploidia e evolução vegetal	4
2.3	Poliploidia e melhoramento de plantas	6
2.4	Aplicações	7
2.5	Verificando níveis de poliploidia	7
3	Modelagem e Simulação	9
3.1	Considerações preliminares	9
3.2	Urnas e bolas, uma caracterização de genotipagem.	13
3.2.1	O modelo de uma urna	13
3.2.2	Várias urnas	16
3.3	Introduzindo erros experimentais	17
3.4	Um modelo de massas moleculares para marcadores genéticos	19
3.4.1	Marcadores dialélicos e SNPs	19
3.4.2	Marcadores genéticos multialélicos	21
3.5	Simulações	21
3.5.1	Experimentos	25
3.5.1.1	Amostra aleatória simples	25
3.5.1.2	Cruzamentos F1	31
4	Genotipagem e Estimação de Ploidia	36
4.1	Eliminação de confundimento	36
4.2	Classificadores	39
4.2.1	Caso dialélico	40
4.2.2	Caso multialélico	40
4.3	Estimação de ploidia - seleção de modelos	42
4.3.1	Medidas de distância interna de agrupamentos	42
4.3.2	Medidas repetidas	43
4.3.3	Medidas de distância entre de agrupamentos	44
4.3.4	Estimadores de ploidia	46
4.4	Genotipagem	46
4.4.1	Estimativa de genotipagem correta	46
4.5	Desenhos experimentais e o Equilíbrio de Hardy-Weinberg	49
4.5.1	Amostra Aleatória Simples (AAS)	49
4.5.2	Experimentos F1	50
4.6	Análise de Dados Simulados	50
4.6.1	Cruzamentos F1	51

4.6.2	Amostra aleatória simples	53
4.6.3	Consistência Empírica	56
4.7	Fluxo da Análise	57
5	Aplicações	58
5.1	Genotipagem de cana-de-açúcar	58
5.1.1	A cana-de-açúcar	58
5.2	Resultados da análise	58
5.2.1	SNP 4	61
5.2.2	SNP 5	65
5.2.3	SNP 13	69
5.2.4	SNP 37	73
5.2.5	SNP 41	77
5.2.6	SNP 45	81
5.2.7	SNP 48	85
5.2.8	SNP 50	89
5.2.9	SNP 60	93
5.2.10	SNP 65	97
5.2.11	SNP 77	101
5.2.12	SNP 79	105
5.2.13	SNP 82	109
5.2.14	SNP 88	113
5.2.15	SNP 114	117
5.2.16	SNP 117	121
5.2.17	SNP 122	125
5.2.18	SNP 136	129
5.2.19	SNP 151	133
5.2.20	SNP 162	137
5.2.21	SNP 186	141
5.2.22	SNP 201	145
5.2.23	SNP 204	149
5.2.24	SNP 225	153
5.2.25	SNP 235	157
5.2.26	SNP 237	161
5.2.27	SNP 241	165
6	Considerações finais e próximos passos	169
6.1	Comparação com o método SuperMASSA	169
6.2	Contribuições	169
6.3	Limitações	171
6.4	Oportunidades	171
A	Códigos R desenvolvidos	176
A.1	graficos.r	176
A.2	polyploid.r	180
A.3	simulation.r	196
A.4	experimentoF1.r	200
A.5	experimentoAAS.r	203
A.6	report.r	206
A.7	sugarcane2.r	212

B Full Bayesian Significance Test (FBST)	216
B.1 Algumas definições	216
B.2 Consistência e Níveis de Significância	217
B.3 Procedimento para o teste e FBST Ilustrado	218
B.3.1 Teste de Proporção	218
B.3.2 Teste de Homogeneidade	219
B.3.3 Modelo Normal	220
C O Equilíbrio de Hardy-Weinberg	225
C.1 Equilíbrio de Hardy-Weinberg	225
C.1.1 Caso Multialélico Diplóide	226
C.1.2 Caso Dialélico Poliplóide	228
C.1.3 Caso Multialélico Poliplóide	229
D Plataforma Affymetrix	230
E Pré-processamento	234
E.1 Remoção de Fundo	234
E.2 Normalização	235
E.3 Genotipagem de SNPs em Humanos	236
F Agrupamentos	238
G Remoção de Fundo	239
G.1 Modelo Normal-Exponencial	239
G.2 Estimação dos parâmetros do modelo Normal-Exponencial:	242
H Métodos de Normalização	243
H.1 Lowess	243
H.2 Normalização quantílica	244
I Técnicas de Agrupamento	246
I.1 Medidas de Distância	246
I.2 Métodos de União	246
I.3 Agrupamento Hierárquico	247
I.4 Agrupamento k-médias	247

Lista de Figuras

1.1	A molécula de DNA 1 difere da molécula de DNA 2 em apenas um par de base do fragmento (polimorfismo C/T)	1
1.2	Representação dos haplótipos de genomas: haplóide (1 cópia), diplóide (2 cópias), triplóide (3 cópias) e tetraplóide (4 cópias) e hexaplóide (6 cópias)	2
2.1	Célula diplóide que apresenta falha na meiose gerando dois gametas diplóides e se autofertilizam produzindo um indivíduo tetraplóide.	5
2.2	Exemplos de culturas poliplóides	6
3.1	A molécula de DNA 1 difere da molécula de DNA 2 em apenas um par de base do fragmento (polimorfismo C/T)	10
3.2	Exemplo de um padrão observado em experimentos de SNP MassArray.	11
3.3	Dados reais de um experimento com cana-de-açúcar.	12
3.4	Ilustração de um genótipo para um determinado loco genético.	13
3.5	O espectômetro de massa é análogo a uma balança capaz de medir separadamente a massa das bolas do tipo A e do tipo B contidas em uma urna.	13
3.6	Possíveis configurações de massas para urnas de capacidade 2, 3, 4 e 5 bolas.	15
3.7	Modelo de várias urnas com capacidades variando de 2 a 5 bolas.	16
3.8	Simulação de modelos de uma urna com capacidades variando de 2 a 5 bolas.	18
3.9	Simulação de modelos de várias urnas com capacidades variando de 2 a 5 bolas.	19
3.10	Dados simulados AAS: Ploidia 2, $N=100\text{mil}$, $p_A = 30\%$, erros com desvios em 0,75; 0,50, 0,25 e 0,10	26
3.11	Dados simulados AAS: Ploidia 2, $N=100\text{mil}$, $p_A = 5\%$, erros com desvios em 0,75; 0,50, 0,25 e 0,10	27
3.12	Dados simulados AAS: Ploidia 6, $N=100\text{mil}$, $p_A = 30\%$, erros com desvios em 0,75; 0,50, 0,25 e 0,10	28
3.13	Dados simulados AAS: Ploidia 6, $N=100\text{mil}$, $p_A = 5\%$, erros com desvios em 0,75; 0,50, 0,25 e 0,10	29
3.14	Dados simulados AAS: Ploidia 12, $N=100\text{mil}$, $p_A = 30\%$, erros com desvios em 0,75; 0,50, 0,25 e 0,10	30
3.15	Dados simulados AAS: Ploidia 12, $N=100\text{mil}$, $p_A = 5\%$, erros com desvios em 0,75; 0,50, 0,25 e 0,10	31
3.16	Dados simulados F1: Ploidia 10, $N=100\text{mil}$, cruzamento 10A x 7A, erros com desvios em 0,75; 0,50, 0,25 e 0,10	32
3.17	Dados simulados F1:Ploidia 10, $N=100\text{mil}$, cruzamento 10A x 5A, erros com desvio em 0,75; 0,50, 0,25 e 0,10	33
3.18	Dados simulados F1:Ploidia 10, $N=100\text{mil}$, cruzamento 10A x 3A, erros com desvios em 0,75; 0,50, 0,25 e 0,10	34
3.19	Dados simulados F1:Ploidia 10, $N=100\text{mil}$, cruzamento 8A x 3A, erros com desvios em 0,75; 0,50, 0,25 e 0,10	35
4.1	Combinação convexa no caso dialélico.	36

4.2	Histogramas de dados transformados pela transformação de projeção na semireta combinação convexa $(0, 1]$, aplicando $U = \frac{X}{X+Y}$, utilizando os dados da Figura 3.9.	38
4.3	Histogramas de dados transformados pela transformação de projeção na combinação convexa $[0;p]$ aplicando $U = p \frac{X}{X+Y}$, utilizando os dados da Figura 3.9.	39
4.4	Combinação convexa no caso trialélico. O semiplano triangular representa a restrição do modelo de pesos moleculares.	41
4.5	Regiões de classificação (genotipagem) nos casos trialélico diplóide e triplóide.	42
4.6	Regiões de classificação (genotipagem) com dados reais de cana-de-açúcar.	43
4.7	Densidades nos pontos de classificação para diferentes ploidias candidatas	45
5.1	Representação do genoma de um cultivar de cana moderno.	59
5.2	Gráficos descritivos dos dos picos e áreas do espectro de massa: histogramas, distribuições acumuladas empíricas. Os dados dos pais estão em destaque.	61
5.3	Gráficos de dispersão e genotipagem para as diferentes medidas aplicadas.	62
5.4	Gráficos descritivos dos dos picos e áreas do espectro de massa: histogramas, distribuições acumuladas empíricas. Os dados dos pais estão em destaque.	65
5.5	Gráficos de dispersão e genotipagem para as diferentes medidas aplicadas.	66
5.6	Gráficos descritivos dos dos picos e áreas do espectro de massa: histogramas, distribuições acumuladas empíricas. Os dados dos pais estão em destaque.	69
5.7	Gráficos de dispersão e genotipagem para as diferentes medidas aplicadas.	70
5.8	Gráficos descritivos dos dos picos e áreas do espectro de massa: histogramas, distribuições acumuladas empíricas. Os dados dos pais estão em destaque.	73
5.9	Gráficos de dispersão e genotipagem para as diferentes medidas aplicadas.	74
5.10	Gráficos descritivos dos dos picos e áreas do espectro de massa: histogramas, distribuições acumuladas empíricas. Os dados dos pais estão em destaque.	77
5.11	Gráficos de dispersão e genotipagem para as diferentes medidas aplicadas.	78
5.12	Gráficos descritivos dos dos picos e áreas do espectro de massa: histogramas, distribuições acumuladas empíricas. Os dados dos pais estão em destaque.	81
5.13	Gráficos de dispersão e genotipagem para as diferentes medidas aplicadas.	82
5.14	Gráficos descritivos dos dos picos e áreas do espectro de massa: histogramas, distribuições acumuladas empíricas. Os dados dos pais estão em destaque.	85
5.15	Gráficos de dispersão e genotipagem para as diferentes medidas aplicadas.	86
5.16	Gráficos descritivos dos dos picos e áreas do espectro de massa: histogramas, distribuições acumuladas empíricas. Os dados dos pais estão em destaque.	89
5.17	Gráficos de dispersão e genotipagem para as diferentes medidas aplicadas.	90
5.18	Gráficos descritivos dos dos picos e áreas do espectro de massa: histogramas, distribuições acumuladas empíricas. Os dados dos pais estão em destaque.	93
5.19	Gráficos de dispersão e genotipagem para as diferentes medidas aplicadas.	94
5.20	Gráficos descritivos dos dos picos e áreas do espectro de massa: histogramas, distribuições acumuladas empíricas. Os dados dos pais estão em destaque.	97
5.21	Gráficos de dispersão e genotipagem para as diferentes medidas aplicadas.	98
5.22	Gráficos descritivos dos dos picos e áreas do espectro de massa: histogramas, distribuições acumuladas empíricas. Os dados dos pais estão em destaque.	101
5.23	Gráficos de dispersão e genotipagem para as diferentes medidas aplicadas.	102
5.24	Gráficos descritivos dos dos picos e áreas do espectro de massa: histogramas, distribuições acumuladas empíricas. Os dados dos pais estão em destaque.	105
5.25	Gráficos de dispersão e genotipagem para as diferentes medidas aplicadas.	106
5.26	Gráficos descritivos dos dos picos e áreas do espectro de massa: histogramas, distribuições acumuladas empíricas. Os dados dos pais estão em destaque.	109
5.27	Gráficos de dispersão e genotipagem para as diferentes medidas aplicadas.	110
5.28	Gráficos descritivos dos dos picos e áreas do espectro de massa: histogramas, distribuições acumuladas empíricas. Os dados dos pais estão em destaque.	113
5.29	Gráficos de dispersão e genotipagem para as diferentes medidas aplicadas.	114

5.30	Gráficos descritivos dos dos picos e áreas do espectro de massa: histogramas, distribuições acumuladas empíricas. Os dados dos pais estão em destaque.	117
5.31	Gráficos de dispersão e genotipagem para as diferentes medidas aplicadas.	118
5.32	Gráficos descritivos dos dos picos e áreas do espectro de massa: histogramas, distribuições acumuladas empíricas. Os dados dos pais estão em destaque.	121
5.33	Gráficos de dispersão e genotipagem para as diferentes medidas aplicadas.	122
5.34	Gráficos descritivos dos dos picos e áreas do espectro de massa: histogramas, distribuições acumuladas empíricas. Os dados dos pais estão em destaque.	125
5.35	Gráficos de dispersão e genotipagem para as diferentes medidas aplicadas.	126
5.36	Gráficos descritivos dos dos picos e áreas do espectro de massa: histogramas, distribuições acumuladas empíricas. Os dados dos pais estão em destaque.	129
5.37	Gráficos de dispersão e genotipagem para as diferentes medidas aplicadas.	130
5.38	Gráficos descritivos dos dos picos e áreas do espectro de massa: histogramas, distribuições acumuladas empíricas. Os dados dos pais estão em destaque.	133
5.39	Gráficos de dispersão e genotipagem para as diferentes medidas aplicadas.	134
5.40	Gráficos descritivos dos dos picos e áreas do espectro de massa: histogramas, distribuições acumuladas empíricas. Os dados dos pais estão em destaque.	137
5.41	Gráficos de dispersão e genotipagem para as diferentes medidas aplicadas.	138
5.42	Gráficos descritivos dos dos picos e áreas do espectro de massa: histogramas, distribuições acumuladas empíricas. Os dados dos pais estão em destaque.	141
5.43	Gráficos de dispersão e genotipagem para as diferentes medidas aplicadas.	142
5.44	Gráficos descritivos dos dos picos e áreas do espectro de massa: histogramas, distribuições acumuladas empíricas. Os dados dos pais estão em destaque.	145
5.45	Gráficos de dispersão e genotipagem para as diferentes medidas aplicadas.	146
5.46	Gráficos descritivos dos dos picos e áreas do espectro de massa: histogramas, distribuições acumuladas empíricas. Os dados dos pais estão em destaque.	149
5.47	Gráficos de dispersão e genotipagem para as diferentes medidas aplicadas.	150
5.48	Gráficos descritivos dos dos picos e áreas do espectro de massa: histogramas, distribuições acumuladas empíricas. Os dados dos pais estão em destaque.	153
5.49	Gráficos de dispersão e genotipagem para as diferentes medidas aplicadas.	154
5.50	Gráficos descritivos dos dos picos e áreas do espectro de massa: histogramas, distribuições acumuladas empíricas. Os dados dos pais estão em destaque.	157
5.51	Gráficos de dispersão e genotipagem para as diferentes medidas aplicadas.	158
5.52	Gráficos descritivos dos dos picos e áreas do espectro de massa: histogramas, distribuições acumuladas empíricas. Os dados dos pais estão em destaque.	161
5.53	Gráficos de dispersão e genotipagem para as diferentes medidas aplicadas.	162
5.54	Gráficos descritivos dos dos picos e áreas do espectro de massa: histogramas, distribuições acumuladas empíricas. Os dados dos pais estão em destaque.	165
5.55	Gráficos de dispersão e genotipagem para as diferentes medidas aplicadas.	166
B.1	FBST para proporção, $(x, y) = (12, 24)$	219
B.2	Visão do plano (π_1, π_2) para o teste de homogeneidade	220
B.3	Densidade posteriori de $f(\mu, \tau)$: Curvas das Hipóteses 0, 1 e 2	221
B.4	Cortando a posteriori para definir o conjunto T_0	222
B.5	Cortando a densidade posteriori para definir os conjuntos T_0 e T_1	222
B.6	Cortando a densidade posteriori para definir os conjuntos T_0 e T_1 e T_2	223
B.7	Curvas de Nível: Conjuntos Tangentes T_0, T_1, T_2, T_3, T_4 , e T_5	223
D.2	GeneChip da Affymetrix	230
D.1	Processo de obtenção da imagem dos chips	231

D.3	Fenômeno de hibridização das moléculas no chip	231
D.4	Fenômeno de hibridização das moléculas no chip	232
D.5	Fenômeno de hibridização das moléculas no chip	232
D.6	Exemplo de dados de SNPs para o caso diplóide, o genótipo heterozigoto AB forma um agrupamento em torno da reta que demarca níveis de expressão iguais em cada alelo.	233
E.1	O sinal observado em um experimento de <i>microarray</i> é composto por um ruído experimental somado ao sinal biológico	235
E.2	Gráficos MA-plot	236
H.1	Antes e depois da normalização Lowess	244
H.2	Antes e depois da normalização quantílica	245

Lista de Tabelas

1.1	Possíveis genótipos para SNPs com haplótipos alelos A e B em função da ploidia de seu locus.	2
3.1	Possíveis configurações de massas em urnas com capacidade fixada.	14
4.1	Exemplo de genotipagem para uma a observação $(A, B) = (0, 64; 0, 21)$. O classificador $h_{2,p}$ (“inteiro mais próximo”) é apresentado na coluna a , que representa o número de possíveis alelos A na observação. O número de alelos b é obtido pela equação da combinação convexa $p = a + b$.	40
4.2	Análise SNP 225 - picos	44
4.3	Probabilidades condicionais de genotipagem correta com hipótese de erros nas vizinhanças do verdadeiro genótipo.	47
4.4	Probabilidades condicionais de genotipagem correta com hipótese de simetria.	48
4.5	Proporções esperadas de genótipos em função da proporção populacional de alelos	49
4.6	Proporções esperadas de genótipos em função da proporção populacional de alelos	49
4.7	Proporções genotípicas esperadas no caso diplóide.	50
4.8	Proporções genotípicas esperadas no caso triplóide.	50
4.9	Número de acertos na estimação de ploidia em simulação de experimentos F1	51
4.10	Proporção de acertos na simulação de experimentos F1	51
4.11	Proporção de acertos na simulação de experimentos F1	52
4.12	Proporção de acertos na simulação de experimentos F1	52
4.13	Proporção de acertos na simulação de experimentos F1	52
4.14	Número de acertos na estimação de ploidia em simulação de experimentos AAS	53
4.15	Proporção de acertos de genotipagem nas simulações de experimentos AAS	54
4.16	Proporção de acertos de genotipagem nas simulações de experimentos AAS	55
4.17	Proporção de acertos de genotipagem nas simulações de experimentos AAS	56
5.1	Análise SNP 4 - picos	63
5.2	Análise SNP 4 - areas	64
5.3	Análise SNP 5 - picos	67
5.4	Análise SNP 5 - areas	68
5.5	Análise SNP 13 - picos	71
5.6	Análise SNP 13 - areas	72
5.7	Análise SNP 37 - picos	75
5.8	Análise SNP 37 - areas	76
5.9	Análise SNP 41 - picos	79
5.10	Análise SNP 41 - areas	80
5.11	Análise SNP 45 - picos	83
5.12	Análise SNP 45 - areas	84
5.13	Análise SNP 48 - picos	87
5.14	Análise SNP 48 - areas	88
5.15	Análise SNP 50 - picos	91
5.16	Análise SNP 50 - areas	92
5.17	Análise SNP 60 - picos	95
5.18	Análise SNP 60 - areas	96

5.19	Análise SNP 65 - picos	99
5.20	Análise SNP 65 - areas	100
5.21	Análise SNP 77 - picos	103
5.22	Análise SNP 77 - areas	104
5.23	Análise SNP 79 - picos	107
5.24	Análise SNP 79 - areas	108
5.25	Análise SNP 82 - picos	111
5.26	Análise SNP 82 - areas	112
5.27	Análise SNP 88 - picos	115
5.28	Análise SNP 88 - areas	116
5.29	Análise SNP 114 - picos	119
5.30	Análise SNP 114 - areas	120
5.31	Análise SNP 117 - picos	123
5.32	Análise SNP 117 - areas	124
5.33	Análise SNP 122 - picos	127
5.34	Análise SNP 122 - areas	128
5.35	Análise SNP 136 - picos	131
5.36	Análise SNP 136 - areas	132
5.37	Análise SNP 151 - picos	135
5.38	Análise SNP 151 - areas	136
5.39	Análise SNP 162 - picos	139
5.40	Análise SNP 162 - areas	140
5.41	Análise SNP 186 - picos	143
5.42	Análise SNP 186 - areas	144
5.43	Análise SNP 201 - picos	147
5.44	Análise SNP 201 - areas	148
5.45	Análise SNP 204 - picos	151
5.46	Análise SNP 204 - areas	152
5.47	Análise SNP 225 - picos	155
5.48	Análise SNP 225 - areas	156
5.49	Análise SNP 235 - picos	159
5.50	Análise SNP 235 - areas	160
5.51	Análise SNP 237 - picos	163
5.52	Análise SNP 237 - areas	164
5.53	Análise SNP 241 - picos	167
5.54	Análise SNP 241 - areas	168
6.1	Comparação de métodos SuperMASSA vs. Método de Urnas.	170
B.1	Valores Críticos de $\overline{ev}$	218
B.2	Evidências contra hipóteses do modelo	221
B.3	Pontos de máxima posteriori sob as hipóteses do modelo	224
C.1	Frequências genotípicas esperadas no Equilíbrio de Hardy-Weinberg	225

Capítulo 1

Introdução

1.1 Considerações preliminares

Um *SNP* (*single nucleotide polymorphism*) é uma variação de um fragmento de *DNA* que ocorre quando, em um determinado loco cromossômico (posição na cadeia cromossômica), apenas um nucleotídeo – A, T, C ou G – difere entre membros de uma espécie ou em pares de cromossomos do mesmo indivíduo. Por exemplo, as sequências AACTAAC e AATTAAC representadas na Figura 1.1, que se encontram na mesma posição de um genoma de indivíduos diferentes, diferem em apenas uma posição de pares de base. Neste caso podemos dizer que existem dois alelos: C e T (com as respectivas bases complementares G e A).

Figura 1.1: A molécula de DNA 1 difere da molécula de DNA 2 em apenas um par de base do fragmento (polimorfismo C/T)

Em organismos diplóides (como os seres humanos), existem duas (não exatamente iguais) “cópias” de cada cromossomo (homólogos), um fragmento de cromossomo individualmente é denominado *haplótipo*, quando nos referimos à descrição do par, a esta descrição chamamos de *genótipo*. Os SNPs compreendem a variação genética mais comum de uma espécie e portanto, são os responsáveis por grande parte variedade de *haplótipos* e *genótipos*. Frequentemente os SNPs possuem apenas dois alelos, que são os mais comuns encontrados em uma determinada população.

No caso de plantas, os cromossomos não necessariamente estão organizados em pares, sua ploidia (número

de cromossomos homólogos) pode variar ao longo do genoma.

Figura 1.2: Representação dos haplótipos de genomas: haplóide (1 cópia), diplóide (2 cópias), triplóide (3 cópias) e tetraplóide (4 cópias) e hexaplóide (6 cópias)

Exemplificamos abaixo os possíveis genótipos para um SNP em função de sua ploidia:

Ploidia	Genótipos
2	AA
	AB
	BB
3	AAA
	AAB
	ABB
	BBB
4	AAAA
	AAAB
	AABB
	ABBB
	BBBB
⋮	⋮

Tabela 1.1: Possíveis genótipos para SNPs com haplótipos alelos A e B em função da ploidia de seu locus.

No mapeamento de genes em doenças ou fenótipos complexos, as plataformas de marcadores moleculares denominadas de *chips* de informação têm trazido novos desafios aos estatísticos, desde a fase de planejamento do experimento até a coleta dos dados e sua análise. Por exemplo, a plataforma de SNPs da *Affymetrix 6.0* (ver Apêndice D) para pares de genes em doenças humanas é muito sensível a fontes de variação devido à técnica envolvida e à qualidade do DNA extraído, o que tem motivado a proposição de novos métodos de processamento e normalização de dados (ver) (Ziegler et al., 2008)[Ziegler2008]. Também, nesses chips a genotipagem dos locos em termos da constituição alélica não é obtida diretamente o que tem incentivado o desenvolvimento de algoritmos de agrupamento específicos ([Carvalho2006] Carvalho et al., 2006).

Plataformas de SNPs para amostragem do genoma têm sido amplamente adotadas em estudos com populações humanas, considerando tanto delineamentos com indivíduos não relacionados (estudos caso-controle) como com indivíduos relacionados (estudos com famílias/pedigrees). O sucesso de tais estudos em termos da identificação de genes depende fortemente do uso apropriado de metodologias estatísticas de análise de dados. Se o cenário deste tipo de pesquisa em populações humanas requer cuidados e envolve desafios, com maior razão esta é a realidade para a utilização de plataformas de SNPs em estudos genômicos de outras populações de animais e plantas. Vale notar, que mapas de SNPs ainda estão sendo construídos para muitas espécies de plantas de importância econômica, como milho e cana de açúcar, por exemplo, pois a ocorrência deste tipo de polimorfismo no genoma destes indivíduos ainda não é completamente conhecida. A poliploidia por sua vez traz desafios ainda mais pertinentes e inviabiliza a aplicação direta de ferramentas de geração e análise de dados já disponíveis para organismos diplóides.

1.2 Desafios e Objetivos

Como veremos no desenvolvimento deste trabalho, os dados provenientes de experimentos de SNPs na plataforma massArray e mais genericamente, experimentos envolvendo pesos moleculares de marcadores genéticos, possuem determinados padrões, que levantam as seguintes questões:

1. É possível simular experimentos que tratam de pesos moleculares de marcadores genéticos? Como?
2. Como genotipar amostras de organismos com ploidia desconhecida?

O objetivo deste trabalho é o de desenvolver:

1. Um modelo estatístico que descreva dados de experimentos com pesos moleculares de marcadores genéticos, permitindo assim, simulações. Detalhamos isso no Capítulo 3.
2. Um critério de genotipagem (ou um classificador) flexível para lidar com amostras de organismos poliplóides. A proposta desenvolvida encontra-se no Capítulo 4.

Uma observação importante: neste trabalho consideramos a ploidia de um loco genético, como sendo uma característica da espécie e não do indivíduo, isto é, não supomos que haja diferença de ploidia entre indivíduos da mesma espécie, progenitores possuem a mesma ploidia em um loco, e ela é preservada na geração seguinte.

1.3 Organização do Trabalho

O Capítulo 2 aborda o fenômeno biológico da poliploidia, sua importância, características, origens e papel no processo evolucionário dos vegetais. No Capítulo ?? trata sobre experimentos com marcadores de SNPs em especial os da plataforma massArray da Sequenom, buscando entender como são obtidos os dados para modelagem e análise estatística. O Capítulo 3 caracteriza o problema de genotipagem de poliplóides em experimentos com SNPs, com o objetivo de construir um modelo adequado para simular padrões observados em dados experimentais com SNPs na plataforma massArray. No Capítulo 4 a meta é elaborar um classificador que genotipe amostras poliplóides levando em consideração o desconhecimento a priori da ploidia dos indivíduos observados. O Capítulo 5 apresenta aplicações das técnicas desenvolvidas em dados simulados e reais, provenientes de estudos de cana-de-açúcar.

Capítulo 2

Poliploidia

O objetivo deste capítulo é o de entender o fenômeno biológico da poliploidia, sua origem, ocorrências e importância no processo evolutivo das plantas. A leitura deste capítulo é opcional para compreensão do modelo estatístico apresentado como objetivo final deste trabalho. Este capítulo baseia-se no trabalho de Thomas G. Ranney [Ranney2000].

2.1 Definição e origens

Um poliplóide é simplesmente um organismo que contém mais do que os habituais dois conjuntos de cromossomos. Nos animais, esta é uma ocorrência bastante rara. Nas plantas, no entanto, isso é muito comum e tem desempenhado um papel evolucionário importante. O desenvolvimento de poliplóides também pode ser uma abordagem útil e valiosa em programas de melhoramento genético de plantas.

O termo "ploidia" ou "nível de ploidia" se refere ao número de conjuntos de cromossomos e é simbolizada por um "n". Um indivíduo com dois conjuntos de cromossomos é referido como um diplóide ($2n$), três conjuntos seria um triplóide ($3n$), e assim por diante.

A poliploidia pode surgir naturalmente de diferentes maneiras. Em alguns casos, uma mutação somática (não reprodutiva) pode ocorrer, devido a uma ruptura na mitose, resultando em duplicação cromossômica que irá subsequentemente produzir um indivíduo poliplóide. Poliplóides também podem resultar da união de gametas não reduzidos - óvulos e espermatozóides que não foram submetidos a meiose normal como esquematizado na Fig 2.1.

A origem de um poliplóide muitas vezes pode determinar sua fertilidade e pode ainda indicar como utilizá-lo em um programa de melhoramento de plantas. Se um tetraplóide surge da duplicação espontânea ou da união de gametas não reduzidos a partir de dois indivíduos diplóides, apesar de diferentes origens, ambos comportam-se similarmente na reprodução e são referidos como autopoliplóides. Autopoliplóides podem ou não ser férteis. Nos diplóides, a meiose envolve o emparelhamento de cromossomos homólogos que eventualmente são separados para formar dois gametas distintos, cada um com um conjunto de cromossomos. A infertilidade pode surgir em autopoliplóides devido ao fato de haver mais de dois cromossomos homólogos. A presença de múltiplos cromossomos homólogos muitas vezes resulta em emparelhamentos espúrios, cromossomos não pareados, e gametas com números cromossômicos desequilibrados (aneuplóides).

Indivíduos resultantes de reprodução sexual com gametas não reduzidos de diferentes espécies são referidos como alopólóides. Devido a sua composição cromossômica, alopólóides são tipicamente férteis. Durante a meiose cada cromossomo pode emparelhar com o seu parceiro homólogo, resultando em células germinativas férteis. Várias culturas de cana-de-açúcar cultivadas para fins comerciais possuem esta característica.

2.2 Poliploidia e evolução vegetal

Em contraste com o processo evolutivo gradual por meio do qual novas espécies evoluem a partir de populações isoladas, novas espécies de plantas podem aparecer abruptamente. O mecanismo mais comum para especiação abrupta é por meio da formação de poliplóides naturais. Por exemplo, uma vez que um tetraplóide surge numa

Figura 2.1: Célula diplóide que apresenta falha na meiose gerando dois gametas diplóides e se autofertilizam produzindo um indivíduo tetraplóide.

população, ele pode hibridar com outros tetraplóides. No entanto, estes tetraplóides são reprodutivamente isolados de suas espécies parentais. Tetraplóides que cruzam com diplóides das espécies parentais resultam em triploídes que são tipicamente estéreis. Esse fenômeno gera uma "barreira reprodutiva" entre os poliplóides e as espécies parentais - uma força motriz para a especiação.

Há uma série de fatores que podem fornecer poliplóides com vantagens adaptativas e evolutivas. Talvez o mais importante seja o fato de poliplóides poderem ter significativamente mais heterozigotos do que suas contrapartes diplóides. O grau de heterozigose pode ser o fator chave no crescimento, desempenho e capacidade de adaptação de uma espécie. Alopoliplóides podem ter um grau muito maior de heterozigoidade, o que pode contribuir para o vigor do híbrido.

Figura 2.2: Exemplos de culturas poliplóides

Uma vez que todos poliplóides têm uma certa quantidade de redundância genética, cópias extras de genes podem mutar e divergir resultando novas características, sem comprometer as funções essenciais. Populações poliplóides frequentemente demonstram rearranjo genômico extenso incluindo a origem de novas regiões de DNA ([Arnold, 1997]; [Song et al, 1995]).

Outra questão frequentemente debatida é a da tolerância ao stress. Observa-se um número desproporcional de poliplóides em regiões frias e secas. Alguns argumentam que esta é uma correlação espúria ([Sanford, 1983]) ou, possivelmente, o resultado da mistura de espécies e formação de alopoliplóides durante o período glacial (Stebbins, 1984)[Stebbins1994]. No entanto, eles também podem ter determinadas características com benesses adaptativas. Estudos moleculares têm demonstrado que alopoliplóides exibem "multiplicidade enzimática" (Soltis e Soltis, 1993)[Soltis1993]. Uma vez que alopoliplóides representam uma fusão de dois genomas distintos, estes poliplóides pode potencialmente produzir todas as enzimas produzidas por cada um dos pais, bem como novas enzimas híbridas. Essa multiplicidade enzimática pode fornecer plantas com maior flexibilidade bioquímica, possivelmente ampliando a gama de ambientes em que podem desenvolver-se (Roose e Gottlieb, 1976)[Roose1976].

2.3 Poliploidia e melhoramento de plantas

Considerando a importância na evolução das plantas, é compreensível o interesse no desenvolvimento de poliplóides induzidos quando os inibidores mitóticos foram descobertos na década de 1930. No entanto, apesar do fato de poliplóides terem sido desenvolvidos para muitas culturas importantes, estas plantas são quase sempre inferiores à seus progenitores. A duplicação somática não introduz qualquer novo material genético, mas sim produz cópias adicionais de cromossomos existentes. Este DNA extra deve ser replicado em cada divisão celular. Assim, o tamanho da célula ampliada é associada a poliplóides com desequilíbrios anatômicos.

Apesar das desvantagens da indução somática, ela pode ser valiosa num programa de melhoramento genético ao aumentar o grau de heteroziguidade e seleção de características desejáveis. Na maioria dos casos verifica-se que autopoliplóides contribuem pouco para o melhoramento das plantas, a menos que a

heterozigotidade substancial possa ser incorporada (Sanford, 1983)[Sanford, 1983].

A poliploidia pode resultar em uma ampla gama de efeitos sobre as plantas, mas os efeitos específicos variam entre: espécies, grau de heterozigotidade, nível de ploidia, e mecanismos que relacionam-se com: silenciamento de genes, interações, efeitos de doses genéticas e regulação de características específicas.

2.4 Aplicações

Desenvolvimento de culturas estéreis A introdução e circulação de espécies invasoras pode ser uma ameaça significativa para determinados ecossistemas. Desenvolvimento de formas estéreis em culturas importantes em berçário é uma abordagem ideal para abordar este problema. Ao fazer isso, as plantas podem ser cultivadas e utilizadas para paisagismo eliminando virtualmente qualquer possibilidade de reprodução sexual e tornar a planta invasiva. Há um número de métodos disponíveis para o desenvolvimento de plantas estéreis. No entanto, uma das abordagens mais rápidas e de custo eficaz para induzir a esterilidade de uma planta é criando poliplóides. Na maioria dos casos, estas plantas funcionam normalmente com a exceção de reprodução, especificamente a meiose. Em alguns casos duplicando os cromossomos de uma planta individual (autopoliploidia) irá resultar em esterilidade devido a vários cromossomos homólogos e complicações durante a meiose.

Restauração da fertilidade em híbridos Não é incomum híbridos entre diferentes espécies ou gêneros serem estéreis. Isso geralmente ocorre devido a uma falha dos cromossomos no emparelhamento durante a meiose. Ao dobrar os cromossomos de um híbrido, cria-se essencialmente um aloploplóide e, assim, restaura-se sua fertilidade.

Aprimoramento de resistência a pragas e tolerância ao estresse A indução da poliploidia pode ser usada como meio para aumentar a resistência contra pragas. O aumento do número de cromossomos pode melhorar a expressão e concentração de certos metabólitos e produtos químicos naturais de defesa. No entanto, é pouco conhecida a relação entre a dose do gene, silenciamento de genes e expressão de metabólitos secundários. Uma abordagem promissora é a criação de aloploplóides entre plantas com diversos produtos químicos de defesa endógenos. Uma característica valiosa dos aloploplóides, é que os metabólitos secundários das espécies parentais são tipicamente aditivos. Isto quer dizer que, eles produzem todas as enzimas e metabólitos de ambos os progenitores de forma eficaz, combinando as características de resistência a pragas de cada um e, potencialmente, contribuindo para uma forma muito mais ampla, mais horizontal de resistência a pragas. Uma abordagem semelhante pode ter utilidade para incrementar a tolerância a certas tensões ambientais.

Tamanho e vigor reforçados. Embora o tamanho das células alargadas encontradas em alguns poliplóides pode ter efeitos indesejáveis, às vezes podem ser benéficos. Em algumas plantas a poliploidia resulta em um aumento de tamanho significativo. Frutas de macieiras tetraplóides podem ser duas vezes tão grandes quanto a fruta diplóide, embora elas tendem a ser aguadas e disformes. Para as maçãs, os triploplóides provaram ser um meio termo que combina frutos maiores, mantendo boa qualidade e são cultivadas para a produção comercial. Este tipo de alargamento pode ser particularmente desejável para flores ornamentais. Pétalas de flores podem também ser mais espessas e flores podem ser mais duradouras em plantas poliplóides (Kehr, 1996)[Kehr1996].

2.5 Verificando níveis de poliploidia

Morfologicamente, plantas com elevados níveis de ploidia possuem aumento aparente. O aumento da ploidia muitas vezes resulta no aumento do tamanho das células que por sua vez resulta em maior espessura, folhas e flores mais largas e frutas de dimensões maiores. Na triagem de um grande número de plantas, características visuais são frequentemente úteis para a identificação de poliplóides.

Em última análise, no entanto, é necessário contar os cromossomos e definitivamente documentar o nível de ploidia. Para isso algumas técnicas laboratoriais podem ser aplicadas:

Citometria de fluxo é uma técnica laboratorial que conta e avalia características físico-químicas de células e partículas suspensas em uma amostra biológica diluída em um fluído com isótopos marcadores.

Chips de microarray é uma plataforma laboratorial capaz de identificar níveis de expressão gênica de várias amostras individuais diferentes por meio de análise de imagens geradas por reações luminosas de marcadores radioativos.

Espectometria de massa é uma técnica que mede pesos de amostras a nível molecular (espectro de masa), como uma balança. A plataforma *massArray* da *Sequenom* é capaz de pesar amostras extraídas dos núcleos celulares de vários indivíduos.

Neste trabalho damos ênfase à técnica de espectometria de massa, em particular a da plataforma MassArray, descrita com mais detalhes em [Andrell2004, Ehrich2005, Fujisawa2004].

Capítulo 3

Modelagem e Simulação

Neste capítulo caracterizamos o problema de genotipagem de poliplóides em experimentos com SNPs. Aqui o objetivo é a construção de um modelo adequado para simular padrões observados em experimentos com SNPs na plataforma MassArray da Sequenom.

3.1 Considerações preliminares

No Capítulo 2, vimos que a poliploidia, o fenômeno biológico no qual uma espécie possui dois ou mais conjuntos de cromossomos em seus grupos de homologia, é comum em diversas espécies vegetais e pode ter várias definições:

Autopoliplóides: possuem múltiplos cromossomos provenientes de uma mesma espécie;

Alopoliplóides: possuem cromossomos de duas espécies diferentes, produzindo descendentes híbridos que, geralmente, são estéreis;

Aneuplóides: possuem níveis de ploidia variável ao longo de seus grupos de cromossomos homólogos.

Vimos também que a poliploidia pode ser verificada em bancada, com experimentos de contagem cromossômica em procedimentos, lentos, complexos, caros e sucetíveis a diversas condições laboratoriais. Uma situação que é dificultada em estudos de espécies onde a ploidia é desconhecida a priori ou a aneuploidia acontece.

Com o avanço da biotecnologia, como vimos no Apêndice ??, novas técnicas foram criadas visando a aferição de propriedades intrínsecas do genoma no nível molecular por meio de técnicas robustas, rápidas, altamente reprodutíveis e, por consequência, economicamente atraentes. Nesta categoria, surgiram os experimentos de marcadores moleculares de SNPs, como os da plataforma MassArray.

Figura 3.1: A molécula de DNA 1 difere da molécula de DNA 2 em apenas um par de base do fragmento (polimorfismo C/T)

Como mostra a Figura 3.1, SNPs são pequenos fragmentos de DNA numa mesma região (locus) do genoma diferentes entre si por apenas um par de bases. Em experimentos com SNPs na plataforma MassArray (Seção ??) conseguimos medir com precisão as massas moleculares de amostras em diversos arranjos ou desenhos experimentais.

Ao avaliar os dados de um experimento MassArray da Sequenom [Andrell2004], podemos observar padrões radiais semelhantes aos da Figura 3.2. Os eixos representam medidas de massa molecular (picos do espectro de massa) observadas para um SNP que possui dois alelos denominados A e B. Nesta figura cada ponto representa a medida de um único indivíduo que é independente dos demais. Amostras de uma população pentaplóide (ocorrência de até 6 classes genotípicas).

Figura 3.2: Exemplo de um padrão observado em experimentos de SNP MassArray.

Na Figura 3.3 observamos os padrões produzidos num experimento de cruzamentos controlados (F1) de cana-de-açúcar. O espectro de massa é uma função gerada pelo experimento, da qual são extraídas duas medidas resumo principais: o pico e a área abaixo do espectro. Aqui observamos os dados de picos e áreas para 3 SNPs diferentes, os pontos marcados com quadrados e cruzeiros coloridos (24 pontos em cada gráfico) representam as medidas repetidas dos progenitores em um experimento de cruzamentos F1. Os pontos circulares em cinza claro representam as medidas dos filhos deste cruzamento. Dados gentilmente cedidos pela equipe do Departamento de Genética da ESALQ-USP sob a coordenação do Prof. Dr. Antônio Augusto Franco Garcia. Para efeito de referência visual, em cinza claro plotamos a reta $A = B$.

Figura 3.3: Dados reais de um experimento com cana-de-açúcar.

Figura 3.4: Ilustração de um genótipo para um determinado loco genético.

3.2 Urnas e bolas, uma caracterização de genotipagem.

3.2.1 O modelo de uma urna

Considere uma urna lacrada com capacidade limitada, porém desconhecida e, por simplicidade, considere que ela contém bolas de dois tipos: bolas do tipo A e bolas do tipo B. Apesar de não sabermos quantas bolas de cada tipo há na urna, podemos assumir que cada bola tem uma massa w constante independente do seu respectivo tipo (alelos do SNP). O genótipo de um loco genético pode ser caracterizado por uma urna lacrada com bolas de diferentes tipos como ilustrado na Figura 3.4.

Queremos estimar o número de bolas de cada tipo contido na urna e, para isso, temos à disposição uma balança capaz de aferir a massa das bolas do tipo A (P_A) e a massa das bolas do tipo B (P_B) contidas na urna. Aqui admitimos que nossa balança não possui erros de medida associados. O espectômetro de massa pode ser associado a uma balança capaz de medir a massa de uma urna aferindo a massa correspondente às das bolas do tipo A e do tipo B separadamente conforme Figura 3.5.

Figura 3.5: O espectômetro de massa é análogo a uma balança capaz de medir separadamente a massa das bolas do tipo A e do tipo B contidas em uma urna.

Com as condições e hipóteses apresentadas, podemos tabular as possíveis massas encontradas em urnas de diferentes configurações G e com diferentes capacidades (número total de bolas). Fazemos isso na Tabela 3.1 e podemos observar uma relação de anti-monotonicidade entre as massas das bolas, isto é, à medida que aumentamos o número de bolas do tipo A em uma urna com capacidade p fixada, necessariamente, a massa das bolas do tipo B têm que diminuir, pois elas estão ficando em menor número.

Capacidade (p)	Configuração (G)	P_A	P_B
2	AA	$2w$	0
	AB	w	w
	BB	0	$2w$
3	AAA	$3w$	0
	AAB	$2w$	w
	ABB	w	$2w$
	BBB	0	$3w$
4	AAAA	$4w$	0
	AAAB	$3w$	w
	AABB	$2w$	$2w$
	ABBB	w	$3w$
	BBBB	0	$4w$
5	AAAAA	$5w$	0
	AAAAB	$4w$	w
	AAABB	$3w$	$2w$
	AABBB	$2w$	$3w$
	ABBBB	w	$4w$
	BBBBB	0	$5w$
$\vdots$	$\vdots$		

Tabela 3.1: Possíveis configurações de massas em urnas com capacidade fixada.

Sob esta contextualização e assumindo $w = 1$, graficamos na Figura 3.6 as possíveis configurações de massas encontradas na Tabela 3.1, para urnas com capacidades variando de 2 a 5 bolas. Cada ponto representa uma possível configuração de massas (P_A , P_B) obtidas da medida de um indivíduo.

Podemos observar que os pontos seguem o padrão da equação da combinação convexa:

$$\begin{aligned}
 p.w &= P_A(G) + P_B(G) \\
 &= \sum_{g \in G} \mathbf{1}_{\{g=A\}}.w + \sum_{g \in G} \mathbf{1}_{\{g=B\}}.w \\
 &= \sum_{i=1}^p \mathbf{1}_{\{g_i=A\}}.w + \sum_{i=1}^p \mathbf{1}_{\{g_i=B\}}.w \\
 &= aw + bw,
 \end{aligned} \tag{3.1}$$

em que,

- $w \in \mathbb{R}$, $0 < w < \infty$;
- p , a , $b \in \mathbb{N}$ finitos.

onde:

- G : possível configuração (ou genótipo) da urna (loco genético do indivíduo)
- $g_i(G)$: bola da posição i na configuração G
- $a(G) = \sum_{i=1}^p \mathbf{1}_{\{g_i=A\}}$: número de bolas do tipo A (alelos A do SNP) na configuração g da urna;

- $b(G) = \sum_{i=1}^p \mathbf{1}_{\{g_i=B\}}$: número de bolas do tipo B (alelos B do SNP) na configuração g da urna;
- p : capacidade (ploidia) da urna (loco do SNP em questão);
- w : massa da bola (dos alelos do SNP).

Note que P_A e P_B são dependentes da configuração (genótipo) da urna. Consequentemente as variáveis a e b também, porém, a soma das duas variáveis é independente da configuração da urna, e das massas das bolas, dependendo apenas da capacidade da urna, reduzindo assim a:

$$p = a + b. \quad (3.2)$$

A urna refere-se ao genótipo do SNP de um determinado loco em um indivíduo. Assumimos que a capacidade (ploidia) da urna seja um atributo do loco do SNP, isto é, indivíduos diferentes (ou urnas diferentes referentes ao mesmo loco genético) podem ter configurações (genótipos) diferentes, mas não possuem ploidias diferentes no mesmo loco genético.

Figura 3.6: Possíveis configurações de massas para urnas de capacidade 2, 3, 4 e 5 bolas.

3.2.2 Várias urnas

Sob as mesmas condições da Subseção 3.2.1, considere agora, uma caixa com um número inteiro C de urnas, todas com a mesma capacidade e configuração.

Neste caso a equação que relaciona a massa total da caixa com a configuração das urnas passa a ser:

$$pwC = awC + bwC, \quad (3.3)$$

que é equivalente às Equações 3.1 e 3.2.

Na Figura 3.6, cada ponto corresponde à medida $(P_A, P_B) = (aw, bw)$ de uma caixa diferente, cada caixa com uma urna. Ao adicionarmos urnas com a mesma configuração em cada caixa, observamos a formação de raios partindo da origem $(0, 0)$ do gráfico. Na Figura 3.7 observamos alguns dos possíveis pares $(P_A, P_B) = (awC, bwC)$ observados no caso de caixas com várias urnas. A linha pontilhada demarca os pontos que representam caixas com uma única urna ($C = 1$). Fazendo C aumentar (ou seja o número de urnas), avançamos nos eixos P_A e P_B (na direção positiva), e obtemos semiretas paralelas à demarcada na figura. Cada raio agrupa caixas com uma mesma configuração de bolas. Uma caixa ou ponto seria o equivalente a amostra de um indivíduo, como amostras diferentes possuem concentrações de moléculas diferentes, estas concentrações podem ser descritas pelo número C de urnas contidas na amostra.

Figura 3.7: Modelo de várias urnas com capacidades variando de 2 a 5 bolas.

Observamos na Figura 3.7 que o padrão radial observado no exemplo das Figuras 3.2 e 3.3, que contém dados reais, começa a aparecer. Além disso, podemos deduzir que nas figuras com dados reais, pontos (indivíduos) pertencentes a um mesmo raio possuem mesma configuração de bolas nas urnas (genótipo). Veremos na Seção 3.3 que a introdução de erros experimentais às medidas das massas obtidas em cada ponto, fará com que os modelos apresentados nesta seção aproximem-se dos padrões obtidos em conjuntos de dados reais.

3.3 Introduzindo erros experimentais

Nos laboratórios, aprendemos que as grandezas físicas são determinadas experimentalmente por medidas que estão associadas às suas incertezas intrínsecas, decorrentes das características dos aparelhos utilizados no processo de medição. COLOCAR REFERENCIAS. Neste contexto, adaptamos os modelos da Seção 3.2.

Modelo de uma urna Introduzindo um erro ϵ à cada medida de massa w observada ao modelo de uma urna, supondo erros aditivos, podemos escrever o vetor de massas (P_A, P_B) de uma urna da seguinte forma:

$$(P_A, P_B) = \left(\sum_{i=1}^p \mathbf{1}_{\{g_i=A\}}(w + \epsilon_i), \sum_{i=1}^p \mathbf{1}_{\{g_i=B\}}(w + \epsilon_i) \right) \quad (3.4)$$

$$= \left(aw + \sum_{i=1}^a \epsilon_i, bw + \sum_{i=1}^b \epsilon_i \right). \quad (3.5)$$

Obtemos assim, os seguintes padrões observados na Figura 3.8. Para uma amostra de tamanho 100 em cada possível configuração (a, b) , adicionamos erros experimentais simulados $\epsilon_j \sim N(0; 0, 15)$, conforme a Equação 3.5. Note que mesmo adicionando erro experimental às massas de cada bola, as medidas observadas permanecem ao redor da reta $p = a + b$ (Equação 3.1), diferindo ainda do padrão desejado na Figura 3.2. Esta limitação do modelo é corrigida na proposta a seguir.

Figura 3.8: Simulação de modelos de uma urna com capacidades variando de 2 a 5 bolas.

Modelo de várias urnas Imagine várias caixas fechadas contendo urnas lacradas com bolas, cada caixa contendo urnas com a mesma capacidade e configuração g , sendo que caixas diferentes podem ter urnas com configurações diferentes. Denotando por C_i o número de urnas contidas na i -ésima caixa da amostra, obtemos:

$$(P_A, P_B)_i = C_i \left(aw + \sum_{j=1}^a \epsilon_j, bw + \sum_{j=1}^b \epsilon_j \right). \quad (3.6)$$

Sob a Equação 3.6, a Figura 3.9 mostra alguns padrões de dispersão. Para uma amostra de tamanho 100 em cada possível configuração (a, b) , adicionamos erros experimentais simulados $\epsilon_j \sim N(0; 0, 15)$, e depois multiplicamos por um número de urnas $C_i \sim Gama(3, 2)$, conforme a Equação 3.6.

Na Subseção 3.2.2 definimos as variáveis C_i como sendo inteiros positivos, e consideramos que as massas moleculares das bolas (alelos) unitárias ($w = 1$). Porém, podemos generalizar a modelagem de C_i estendendo à distribuições contínuas estritamente positivas como a Gama, por exemplo. Isto pode ser feito sem perda de generalidade pois, a princípio, não temos informação sobre a massa individual dos alelos de cada marcador genético e, além disso, as medidas obtidas são provenientes de amostras com concentrações diferentes de moléculas. Mesmo em medidas repetidas podemos observar concentrações diferentes.

Figura 3.9: Simulação de modelos de várias urnas com capacidades variando de 2 a 5 bolas.

Pela Figura 3.9, observamos que o modelo de várias urnas com bolas e seus respectivos erros de medida associados parece adequado para gerar padrões semelhantes como os observados em experimentos de SNPs na plataforma MassArray (conforme Figuras 3.2 e 3.3). O conceito de concentração das amostras é fundamental para gerar padrões radiais coerentes com os observados experimentalmente.

3.4 Um modelo de massas moleculares para marcadores genéticos

Nesta seção segue uma formalização do modelo construído ao longo deste capítulo, objetivando a simulação de experimentos envolvendo a aferição de massas moleculares de marcadores genéticos.

3.4.1 Marcadores dialélicos e SNPs

Em SNPs ou marcadores dialélicos, temos que considerar apenas duas possibilidades de polimorfismos, isto é, alelos do tipo A ou alelos do tipo B. Para uma amostra biológica de ploidia p , podemos definir o genótipo G como uma sequência das letras A e B de tamanho p , por exemplo, se $p = 5$, poderíamos observar $G = (A, A, A, B, B)$.

Admitindo que cada alelo possua uma distribuição de massas moleculares observados em que:

$$W_A = w_A + \epsilon_A, \quad (3.7)$$

$$W_B = w_B + \epsilon_B, \quad (3.8)$$

com:

- W_i : massa molecular observada do alelo $i \in \{A, B\}$
- w_i : massa molecular intrínseca do alelo i , constante e desconhecida.
- ϵ_i : erro de medida, aleatório dentro do limite da precisão do aparelho.

Escrevendo os massas moleculares de uma amostra com concentração $C > 0$ em função da ploidia p , do genótipo G e seus respectivos alelos g , temos

$$P(C, p, G) = (P_A(C, p, G), P_B(C, p, G)) \quad (3.9)$$

$$= \left(\sum_{g \in G} C \cdot W_A \cdot 1_{\{g=A\}}, \sum_{g \in G} C \cdot W_B \cdot 1_{\{g=B\}} \right) \quad (3.10)$$

$$= C \left(\sum_{g \in G} (w_A + \epsilon_A) \cdot 1_{\{g=A\}}, \sum_{g \in G} (w_B + \epsilon_B) \cdot 1_{\{g=B\}} \right) \quad (3.11)$$

$$= C \sum_{g \in G} ((w_A + \epsilon_A) \cdot 1_{\{g=A\}}, (w_B + \epsilon_B) \cdot 1_{\{g=B\}}) \quad (3.12)$$

$$= C \left((a \cdot w_A, b \cdot w_B) + \left(\sum_{i=1}^a \epsilon_{A,i}, \sum_{i=1}^b \epsilon_{B,i} \right) \right) \quad (3.13)$$

em que:

- a : número de alelos A,
- b : número de alelos B,
- $a + b = p$.

Uma simplificação Podemos considerar que não haja diferenças entre as massas moleculares intrínsecas dos alelos, isto é, que $w_A = w_B$, já que eles possuem o mesmo número de pares de base, o que nos leva a considerar também que ϵ_A e ϵ_B tem a mesma distribuição.

Assim a equação 3.13, incorporando as restrições nos componentes aleatórios ficaria

$$P(C, p, G) = C \left[w(a, b) + \left(\sum_{j=1}^a \epsilon_j, \sum_{j=1}^b \epsilon_j \right) \right] \quad (3.14)$$

$$= C \left(aw + \sum_{j=1}^a \epsilon_j, bw + \sum_{j=1}^b \epsilon_j \right) = C(X_A, X_B). \quad (3.15)$$

Sujeito à restrição $a + b = p$, $(a, b) \in \{0, \dots, p\}^2$.

Chegamos então a uma expressão simples que é função de três variáveis aleatórias estritamente positivas:

- a concentração da amostra C , a qual fisicamente faz sentido modelar com distribuições de probabilidade estritamente positivas;
- as variáveis X_A e X_B : somas finitas de constantes (aw e bw) e variáveis aleatórias ($\sum_{i=1}^a \epsilon_i$ e $\sum_{i=1}^b \epsilon_i$). Podemos denominar as somas $\sum_{i=1}^a \epsilon_i$ e $\sum_{i=1}^b \epsilon_i$ como sendo as amplitudes dos erros de medida associados à precisão do espectômetro de massa. Para modelar as variáveis ϵ_j podemos utilizar distribuições centradas em 0 e com variância finita.

Uniformização de variâncias As variâncias das medidas ($\sum_{i=1}^a \epsilon_i$ e $\sum_{i=1}^b \epsilon_i$) mudam de acordo com o genótipo. Para uniformizar a variância dos erros entre diferentes genótipos, podemos escrever:

$$P(C, p, G) = C[(a, b) + (\epsilon_a, \epsilon_b)] \quad (3.16)$$

onde $\epsilon_j \sim D_{j\epsilon}$ e $C \sim D_C$, com

- $D_{j\epsilon}$ uma distribuição de média 0 e variância finita,
- D_C uma distribuição contínua estritamente positiva.

3.4.2 Marcadores genéticos multialélicos

O modelo dialélico descrito na Subseção 3.4.1 pode ser estendido para marcadores multialélicos, ou seja, com mais de duas possibilidades de polimorfismo. Se um loco genético em questão tem K possíveis alelos, podemos generalizar o modelo da Equação 3.15.

$$P(C, p, K, G) = C \begin{pmatrix} a_1 w + \sum_{i=1}^{a_1} \epsilon_i \\ a_2 w + \sum_{i=1}^{a_2} \epsilon_i \\ \vdots \\ a_K w + \sum_{i=1}^{a_K} \epsilon_i \end{pmatrix}^T = C \begin{pmatrix} X_{a_1} \\ X_{a_2} \\ \vdots \\ X_{a_K} \end{pmatrix}^T. \quad (3.17)$$

Sujeito à restrição $p = a_1 + a_2 + \dots + a_K$, $a_j \in \{0, \dots, p\}$, $j = 1, \dots, K$.

Ou ainda:

$$P(p, K, G) = C \begin{pmatrix} a_1 + \epsilon_{a_1} \\ a_2 + \epsilon_{a_2} \\ \vdots \\ a_K + \epsilon_{a_K} \end{pmatrix}^T = C \begin{pmatrix} X_{a_1} \\ X_{a_2} \\ \vdots \\ X_{a_K} \end{pmatrix}^T. \quad (3.18)$$

Sujeito à restrição $p = X_1 + X_2 + \dots + X_K$, $X_j \in \{0, \dots, p\}$, $j = 1, \dots, K$.

3.5 Simulações

Nesta seção mostramos como simular genótipos, cruzamentos e medidas experimentais de expressão gênica para fins de simulação.

Genótipos

Para o caso k-alélico o genótipo de um poliplóide com ploidia p pode ser descrito como um vetor no formato:

$$G = (g_1, g_2, \dots, g_p)$$

onde $g_i \in \{1, 2, \dots, k\}$

Podemos gerar este vetor definindo uma distribuição de probabilidade dos k alelos. Gerando p alelos, obtemos o genótipo G acima.

Um exemplo de implementação em linguagem R do método acima pode ser descrito abaixo.

```

1 #####
2 # FUNCTION:          DESCRIPTION:
3 # rGenotype          simulates genotypes
4 #####
5
6 rGenotype <-
7   function (n, p = 2, prob = c(0.5, 0.5), ABC=FALSE)

```

```

8 {
9   # A function implemented by Silvio Rodrigues de Faria Junior
10
11   # Description:
12   # Simulates genotype samples from multiallelic locus with ploidy p
13
14   # Arguments:
15   #   n : sample size (integer)
16   #   p : ploidy (integer). Remember, humans and animals are diploid.
17   #   prob : proportion of alleles (real vector).
18   #   ABC : representation with letters instead numbers.
19
20   # Value:
21   # Returns a (n X p) matrix with genotypes, one per line.
22
23   # Note:
24   #
25
26   # FUNCTION:
27
28   res = sample.int(length(prob), size = n*p, replace = TRUE, prob = prob)
29   if (ABC) res = LETTERS[res]
30   if (length(prob) == 2) res = res - 1
31
32   matrix(res, ncol=p)
33 }
34
35 # Examples
36 if (FALSE) {
37   # 5 size samples
38   #####
39   # Tetraploid sample, same alleles proportion.
40   rGenotype(5, 4)
41   #      [,1] [,2] [,3] [,4]
42   # [1,]  0    0    0    0
43   # [2,]  1    0    1    1
44   # [3,]  0    1    1    1
45   # [4,]  1    1    0    0
46   # [5,]  1    1    1    1
47
48   # Tetraploid sample, different alleles proportion.
49   rGenotype(5, 4, prob=c(0.2, 0.8))
50
51   # Tetraploid sample with 3 alleles.
52   rGenotype(5, 4, prob=c(0.2, 0.4, 0.4))
53
54   # Tetraploid sample with 3 alleles: A, B, C and D.
55   rGenotype(10, 4, prob=c(0.2, 0.4, 0.3, .1), ABC = TRUE)
56
57 }
58

```

Cruzamentos

Para simular os cruzamentos de uma população F1 precisamos:

1. Sortear os alelos do gameta do pai;
2. Sortear os alelos do gameta da mãe;
3. Combinar os gametas

Para os casos de ploidia ímpar, fazemos o sorteio na ploidia sucessora e eliminamos o alelo excedente da combinação de gametas. O método pode ser implementado em código R como ilustrado abaixo.

```

1 #####
2 # FUNCTION: DESCRIPTION:
3 # rF1 generates F1 samples
4 #####
5
6 rF1 <-
7   function (n, father, mother)
8   {
9     # A function implemented by Silvio Rodrigues de Faria Junior
10
11     # Description:

```



```

12      # Calculates genotype proportions of a biallelic locus.
13
14      # Arguments:
15      #     n : sample size (integer)
16      # father : father genotype (0 and 1 vector)
17      # mother : mother genotype (0 and 1 vector)
18
19      # Notes:
20      #     father and mother must have the same length.
21
22      # Value:
23      #     Returns a (n X p) matrix with offspring genotypes,
24      #     where p = length(mother)
25
26      # FUNCTION:
27
28      nfather = length(father)
29      nmother = length(mother)
30      if (nfather != nmother)
31          stop("Father and mother length genotypes cannot differ.")
32      ploidia = nmother
33      ployd = ploidia
34      if ((ployd %% 2) != 0)
35          ployd = ployd + 1
36      fa.gametes = combn(father, ployd/2)
37      mo.gametes = combn(mother, ployd/2)
38      Ngametes = ncol(fa.gametes)
39      fa.choosen = floor(runif(n, 1, Ngametes+1))
40      mo.choosen = floor(runif(n, 1, Ngametes+1))
41      A = t(rbind(fa.gametes[,fa.choosen], mo.gametes[, mo.choosen]))
42      t(apply(A, 1, sample, size=ploydia))
43  }
44
45  if (FALSE) {
46
47      # triploid homozygote crossing
48      rFi(10, c(1,1,1), c(0,0,0))
49
50      # tetraploid homozygote crossing
51      rFi(10, c(1,1,1,1), c(0,0,0,0))
52
53      # tetraploid heterozygote crossing
54      rFi(10, c(1,1,0,0), c(1,1,0,0))
55
56  }

```

Medidas de Expressão

Uma vez que sabemos simular cruzamentos AAS e F1, utilizando os modelos desenvolvidos nas Seção 3.4 como o exemplo apresentado a seguir.

```

1  #####
2  # SIMULAÇÃO
3  # AMOSTRA ALEATÓRIA
4  # Ploidias (p): 2, 6 e 12
5  # Proporções Alélicas (pA): 0.30 e 0.05
6  # Desvios (eps): 0.75; 0.50; 0.25 e 0.10
7  #####
8  # Modelo simplificado (variação). Equação (17). Seção 3.4.1.
9  #####
10 set.seed(2501)
11
12 N = 100000
13 Rw = rweibull(N, 10, 40)
14
15 #simulacoes = list()
16 aas.data = list()
17 i = 1
18 for (ploidia in c(2, 6, 12)) {
19     for (pA in c(.30, .05)) {
20         AA = rGenotype(N, ploidia, c(1-pA, pA))
21         a = rowSums(AA)
22         b = ploidia - a
23         for (eps in c(.75, .50, .25, .1)) {
24             EPS = rnorm(N, 0, eps)
25             P = abs(cbind(a + EPS, b - EPS))
26             Pn = P/rowSums(P)

```

```

27     PAB = Rw * Pn
28
29     arquivo = paste('amostraAleatoria-2-', i, '.png', sep='')
30     png(arquivo)
31     plot(hexbin(PAB[,1], PAB[,2], xbins=50,
32               xbnbs = c(0, 55), ybnbs = c(0, 55),
33               #xbnbs = c(0, max(PAB)), ybnbs = c(0, max(PAB)),
34               xlab = expression(P[A]), ylab = expression(P[B])),
35           main = "Amostra Aleatória Simples")
36     dev.off()
37
38     aas.data[[i]] = list(ploidia = ploidia, pA = pA,
39                        eps=eps, PAB = cbind(PAB, a, b))
40     i = i + 1
41   }
42 }
43 }
44
45 save(aas.data, file="amostraAleatoria.2.RData")
46
47 #####
48 # SIMULAÇÃO
49 # CRUZAMENTOS F1
50 # Ploidias (p): 2, 6 e 12
51 # Proporções Alélicas (pA): 0.30 e 0.05
52 # Desvios (eps): 0.75; 0.50; 0.25 e 0.10
53 #####
54 # Modelo de soma de erros. Equação (16). Seção 3.4.1.
55 #####
56 cruzamentos = list()
57 # CRUZAMENTO 10A x 7A
58 cruzamentos[[1]] = list(c(1,1,1,1,1,1,1,1,1,1), c(1,1,1,1,1,1,0,0,0))
59 # CRUZAMENTO 10A x 5A
60 cruzamentos[[2]] = list(c(1,1,1,1,1,1,1,1,1,1), c(1,1,1,1,0,0,0,0,0))
61 # CRUZAMENTO 10A x 3A
62 cruzamentos[[3]] = list(c(1,1,1,1,1,1,1,1,1,1), c(1,1,1,0,0,0,0,0,0))
63 # CRUZAMENTO 8A x 3A
64 cruzamentos[[4]] = list(c(1,1,1,1,1,1,1,1,0,0), c(1,1,1,0,0,0,0,0,0))
65
66 ploidia = 10
67 #simulacoes2 = list()
68 f1.data = list()
69 i = 1
70 for (cr in cruzamentos) {
71   GF = rF1 (N, cr[[1]], cr[[2]])
72   a = rowSums(GF)
73   b = ploidia - a
74   for (eps in c(.75, .50, .25, .1)) {
75     EPS = rnorm(N, 0, eps)
76     P = abs(cbind(a + EPS, b - EPS))
77     Pn = P/rowSums(P)
78     PAB = Rw * Pn
79
80     arquivo = paste('cruzamentosF1-2-', i, '.png', sep='')
81     png(arquivo)
82     plot(hexbin(PAB[,1], PAB[,2], xbins=50,
83               xbnbs = c(0, 55), ybnbs = c(0, 55),
84               #xbnbs = c(0, max(PAB)), ybnbs = c(0, max(PAB)),
85               xlab = expression(P[A]), ylab = expression(P[B])),
86           main = "Cruzamentos F1")
87     dev.off()
88
89
90
91     EPS = rnorm(N, 0, eps)
92     P = abs(cbind(a + EPS, b - EPS))
93     Pn = P/rowSums(P)
94     PABPai = Rw * Pn
95
96     EPS = rnorm(N, 0, eps)
97     P = abs(cbind(a + EPS, b - EPS))
98     Pn = P/rowSums(P)
99     PABMae = Rw * Pn
100
101     f1.data[[i]] = list(cruz = cr, eps=eps, PAB = cbind(PAB, a, b),
102                      PABPai = PABPai, PABMae=PABMae)
103     i = i + 1
104   }
105 }
106
107 save(f1.data, file="cruzamentosF1.2.RData")

```

3.5.1 Experimentos

Nesta seção exibimos algumas simulações desenvolvidas em diferentes desenhos amostrais. O código das simulações pode ser conferido no Apêndice A.

Em cada simulação, foi gerada uma amostra populacional de tamanho 100mil.

Para cada condição de parâmetros, consideramos um desenho experimental com medidas repetidas de tamanho 12, com amostras de tamanhos 180.

Cada condição experimental foi simulada 500 vezes.

3.5.1.1 Amostra aleatória simples

Simulamos amostras aleatórias sob diferentes condições de:

- Ploidia: 2, 6 e 12 alelos
- Proporções alélicas da população: 30% e 5% do alelo mais raro
- Erros: distribuições normais centradas em zero com desvios 0,75; 0,50; 0,25 e 0,10.

As Figuras 3.10, 3.11, 3.12, 3.13, 3.14 e 3.15 ilustram histogramas bidimensionais em cada uma das condições experimentais avaliadas. Quanto mais escuro o hexágono da figura, mais densa a região.

Figura 3.10: Dados simulados AAS: Ploidia 2, $N=100\text{mil}$, $p_A = 30\%$, erros com desvios em 0,75; 0,50, 0,25 e 0,10

Figura 3.11: Dados simulados AAS: Ploidia 2, $N=100\text{mil}$, $p_A = 5\%$, erros com desvios em 0,75; 0,50, 0,25 e 0,10

Figura 3.12: Dados simulados AAS: Ploidia 6, $N=100\text{mil}$, $p_A = 30\%$, erros com desvios em 0,75; 0,50, 0,25 e 0,10

Figura 3.13: Dados simulados AAS: Ploidia 6, $N=100\text{mil}$, $p_A = 5\%$, erros com desvios em 0,75; 0,50, 0,25 e 0,10

Figura 3.14: Dados simulados AAS: Ploidia 12, $N=100\text{mil}$, $p_A = 30\%$, erros com desvios em 0,75; 0,50, 0,25 e 0,10

Figura 3.15: Dados simulados AAS: Ploidia 12, $N=100\text{mil}$, $p_A = 5\%$, erros com desvios em 0,75; 0,50, 0,25 e 0,10

Quando avaliamos amostras de 100mil indivíduos em diferentes condições de proporções alélicas e diferentes condições de variabilidade, podemos observar mais claramente os padrões gerados por estas condições assim como o efeito da variabilidade dos dados na distinção dos grupos.

3.5.1.2 Cruzamentos F1

Simulamos cruzamentos controlados F1 de um loco de ploidia 10 sob diferentes condições de:

- Cruzamentos:
 - 10A x 7A ou $(A,A,A,A,A,A,A,A,A) \times (A,A,A,A,A,A,B,B,B)$
 - 10A x 5A ou $(A,A,A,A,A,A,A,A,A) \times (A,A,A,A,B,B,B,B,B)$

- 10A x 3A ou (A,A,A,A,A,A,A,A,A) x (A,A,A,B,B,B,B,B)
- 8A x 3A ou (A,A,A,A,A,A,A,B,B) x (A,A,A,B,B,B,B,B)

- Erros: distribuições normais centradas em zero com desvios 0,75; 0,50; 0,25 e 0,10.

As Figuras 3.16, 3.17, 3.18 e 3.19 ilustram histogramas bidimensionais em cada uma das condições experimentais avaliadas. Quanto mais escuro o hexágono da figura, mais densa a região.

Figura 3.16: Dados simulados F1: Ploidia 10, N=100mil, cruzamento 10A x 7A, erros com desvios em 0,75; 0,50, 0,25 e 0,10

Figura 3.17: Dados simulados F1:Ploidia 10, $N=100\text{mil}$, cruzamento 10A x 5A, erros com desvio em 0,75; 0,50, 0,25 e 0,10

Figura 3.18: Dados simulados F1:Ploidia 10, $N=100\text{mil}$, cruzamento 10A x 3A, erros com desvios em 0,75; 0,50, 0,25 e 0,10

Figura 3.19: Dados simulados F1:Ploidia 10, $N=100\text{mil}$, cruzamento $8A \times 3A$, erros com desvios em 0,75; 0,50, 0,25 e 0,10

Quando avaliamos amostras de 100mil indivíduos em diferentes condições de cruzamentos F1 e diferentes condições de variabilidade, podemos observar mais claramente os padrões gerados por estas condições assim como o efeito da variabilidade dos dados na distinção dos grupos.

Capítulo 4

Genotipagem e Estimação de Ploidia

Neste capítulo nossa meta é elaborar um classificador que genotipe amostras de poliplóides. Supondo que a ploidia seja uma característica do loco genético em questão, é possível construir um classificador que considere informações sobre a geração ancestral (pais) e leis de equilíbrio genético (Mendel e Hardy-Weinberg) na formação dos genótipos dos filhos.

4.1 Eliminação de confundimento

Observamos na Seção 3.4 que o modelo que descreve pesos moleculares de SNPs, a menos de uma variável de concentração da amostra biológica, tem um padrão geométrico intrinsecamente ligado à combinação convexa, isto é, se possuímos uma amostra com a genes (alelos A) e b genes (alelos B), num organismo de ploidia p , necessariamente temos que:

$$p = a + b, (a, b) \in \{0, \dots, p\}^2. \quad (4.1)$$

A Figura 4.1 ilustra esse resultado. A semireta representa a restrição do modelo de massas moleculares. Os pontos em destaque representam os possíveis genótipos. Num organismo de ploidia p , observamos $p+1$ pontos que representam os genótipos.

Figura 4.1: Combinação convexa no caso dialélico.

Isto nos motiva a considerar, numa possível análise de classificação dos dados, o uso de uma transformação que reduza a dimensão do problema sem perder o poder de discriminação dos dados que possuem este

padrão. Tal transformação pode ser possível se considerarmos a projeção dos pontos observados na semireta *combinação convexa*, representada na Figura 4.1, dada por:

$$1 = x + y, \quad (x, y) \in [0; 1] \times [0; 1]. \quad (4.2)$$

em que,

- x : representa o peso molecular para o alelo A,
- y : representa o peso molecular para o alelo B.

A Equação 4.2 não faz referência à ploidia p , o que faz sentido pois não conhecemos p à priori, e mesmo desconhecido, a semireta gerada por esta equação é paralela à equação $p = a + b$. Como podemos ver na Figura 4.1 as duas transformações são equivalentes por semelhança de triângulos.

Considerando a seguinte transformação:

$$U = \frac{X}{X + Y} \quad (4.3)$$

que projeta os pontos do semiplano positivo $\{(x, y) : x \geq 0, y \geq 0\}$ na semireta 4.2, tem-se:

$$\left. \begin{aligned} X &= C \left(aw + \sum_{j=1}^a \epsilon_j \right) \\ Y &= C \left(bw + \sum_{j=1}^b \epsilon_j \right) \end{aligned} \right\} \Rightarrow U = \frac{C \left(aw + \sum_{i=1}^a \epsilon_i \right)}{C \left(aw + \sum_{i=1}^a \epsilon_i \right) + C \left(bw + \sum_{i=1}^b \epsilon_i \right)} \\ = \frac{aw + \sum_{i=1}^a \epsilon_i}{pw + \sum_{i=1}^p \epsilon_i} \quad (4.4)$$

A Equação 4.4 nos diz que uma vez projetados na semireta combinação convexa (Equação 4.2), os dados do modelo tornam-se independentes do fator de concentração C , diminuindo assim a dimensão do problema, isto é, projetando o semiplano em uma semireta.

Ao aplicar a transformação U nos dados da Figura 3.9 observamos o seguinte comportamento da variável transformada, indicado na Figura 4.2.

Figura 4.2: Histogramas de dados transformados pela transformação de projeção na semireta combinação convexa $(0, 1]$, aplicando $U = \frac{X}{X+Y}$, utilizando os dados da Figura 3.9.

Como podemos observar na Figura 4.2, os dados simulados para gerar a Figura 3.9 apresentam uma distribuição multimodal ao serem projetados na combinação convexa $[0;1]$.

Se soubermos a priori a ploidia p dos dados em questão, aplicando a transformação

$$U_p(X, Y) = p \frac{X}{X + Y}, \quad (4.5)$$

obtemos histogramas como os ilustrados na Figura 4.3.

Figura 4.3: Histogramas de dados transformados pela transformação de projeção na combinação convexa $[0;p]$ aplicando $U = p \frac{X}{X+Y}$, utilizando os dados da Figura 3.9.

A diferença entre as Figuras 4.2 e 4.3 está apenas na escala do eixo U . Isto é, com a transformação 4.5, as modas da distribuição dos dados ficam centradas nos pontos $\{0, \dots, p\}$, indicando assim o número de alelos A de cada observação e, por consequência da Equação 4.1, o número de alelos B também. Como a priori não sabemos a ploidia p da amostra em questão, sabendo como a amostra deve se comportar no caso de p conhecida, podemos propor critérios para avaliar ploidias candidatas para descrever o comportamento de um experimento.

4.2 Classificadores

Nesta seção vamos supor que a ploidia do loco genético que estamos analisando é conhecida, e a partir dos padrões esperados com relação à distribuição multimodal de dados observada na Seção 4.1, propomos classificadores para genotipar amostras de organismos poliplóides.

Nesta seção, denotamos por A e B , as medidas de massa observadas para os respectivos alelos A e B após aplicarmos a transformação U de 4.3, isto é, dadas as observações de massa (P_A, P_B) , obtemos:

$$\begin{cases} A = U(P_A, P_B) = \frac{P_A}{P_A + P_B} \\ B = U(P_B, P_A) = \frac{P_B}{P_A + P_B} \end{cases}$$

4.2.1 Caso dialélico

Os padrões apresentados pelos histogramas de dados transformados na Seção 4.1 levam-nos a propor como classificador a função $H : [0, 1]^2 \rightarrow \{0, \dots, p\}$

$$H_{2,p}(A, B) = (h_{2,p}(U_p(A, B)), p - h_{2,p}(U_p(A, B))) \quad (4.6)$$

com

$$h_{2,p}(u) = \begin{cases} \lceil u \rceil & |u - \lceil u \rceil| < |u - \lfloor u \rfloor| \\ \lfloor u \rfloor & c.c. \end{cases} \quad (4.7)$$

em que:

- $\lceil u \rceil$: inteiro mais próximo maior que u ,
- $\lfloor u \rfloor$: inteiro mais próximo menor que u

Ou seja, conhecendo a priori a ploidia da amostra, um classificador natural seria a função "inteiro mais próximo". Neste contexto, a classificação é perfeita se os erros de medida forem tais que os valores transformados de cada agrupamento tenham uma amplitude no intervalo $(h_{2,p}(u) - 0,5; h_{2,p}(u) + 0,5)$, como observados na Figura 4.3 que apresenta distribuições multimodais com agrupamentos bem separados uns dos outros e com modas centradas nos pontos $\{0, 1, \dots, p\}$.

Como um exemplo prático de genotipagem, a Tabela 4.1 apresenta possíveis genotipagens de uma observação $(A, B) = (0,64; 0,21)$ em diferentes níveis de ploidia.

Ploidia (p)	$p.A$	$p.B$	$U_{2,p}$	a	b
2	1,28	0,42	1,5058	2	0
3	1,92	0,63	2,2588	2	1
4	2,56	0,84	3,0117	3	1
5	3,20	1,05	3,7647	4	1
6	3,84	1,26	4,5176	5	1
7	4,48	1,47	5,2705	5	2

Ploidia (p)	$p.A$	$p.B$	$U_{2,p}$	a	b
8	5,12	1,68	6,0235	6	2
9	5,76	1,89	6,7764	7	2
10	6,40	2,10	7,5294	8	2
11	7,04	2,31	8,2823	8	3
12	7,68	2,52	9,0352	9	3
13	8,32	2,73	9,7882	10	3
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$

Tabela 4.1: Exemplo de genotipagem para uma a observação $(A, B) = (0,64; 0,21)$. O classificador $h_{2,p}$ ("inteiro mais próximo") é apresentado na coluna a , que representa o número de possíveis alelos A na observação. O número de alelos b é obtido pela equação da combinação convexa $p = a + b$.

4.2.2 Caso multialélico

O SNP é um tipo de marcador genético dialélico, sendo considerado apenas dois tipos de alelos (ou dois tipos de bolas) em cada conjunto de homologia, ou podemos dizer um marcador de dimensão 2. Porém, o classificador proposto é flexível o suficiente para considerarmos o problema em dimensões superiores, ou marcadores multialélicos, onde num grupo de homologia podemos ter mais de 2 tipos de alelos (ou bolas).

As mesmas considerações desenvolvidas nas Seções 4.1 e 4.2.1 podem ser extendidas ao caso multialélico de dimensão arbitrária. Nele a combinação convexa se torna:

$$p = a_1 + a_2 + \dots + a_k, \quad a_j \in \{0, \dots, p\}, \quad j = 1, \dots, K.$$

Figura 4.4: Combinação convexa no caso trialélico. O semiplano triangular representa a restrição do modelo de pesos moleculares.

A Figura 4.4 ilustra o caso em que um marcador tem três alelos (A, B e C).

A transformação U equivalente a Equação 4.5 fica:

$$U_{p,K} = \frac{p}{\sum_{j=1}^K x_j} \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_{K-1} \end{pmatrix}, \quad (4.8)$$

e seu correspondente classificador multialélico é dado por:

$$h_p \begin{pmatrix} u_1 \\ u_2 \\ \vdots \\ u_{K-1} \end{pmatrix} = \begin{cases} \lceil u_j \rceil & |u_j - \lceil u_j \rceil| < |u_j - \lfloor u_j \rfloor|, \\ \lfloor u_j \rfloor & c.c. \end{cases}, \quad j = 1, \dots, K. \quad (4.9)$$

A Figura 4.5 ilustra as regiões de genotipagem nos casos trialélico diplóide ($p = 2$) e triplóide ($p = 3$). A grade hexagonal identifica as regiões em que o classificador falha em atribuir um agrupamento a um indivíduo. Os pontos negros em destaque representam as medidas observadas pelas genotipagens caso não houvesse erro na medição de pesos moleculares.

Figura 4.5: Regiões de classificação (genotipagem) nos casos triploide diplóide e triploide.

4.3 Estimação de ploidia - seleção de modelos

Usualmente, nos problemas envolvendo técnicas de agrupamento (ver Apêndice I) não dispomos de pontos de referência disponíveis a priori para aglomerar os indivíduos da amostra. É necessário o uso de medidas de distância entre indivíduos (que avalie o grau de similaridade interna do grupo), e entre agrupamentos (que avalie o grau de diferenciação entre grupos). Um problema adicional é a estimação do número de agrupamentos caso não haja informação a respeito.

O problema de genotipagem de poliploides é semelhante aos problemas de classificação e agrupamento, exceto pelo fato de que, uma vez definido o número de agrupamentos (ou ploidia) dispomos de pontos de referência em que esperamos uma aglutinação maior dos dados, como visto na Seção 4.1. Uma vez definidos o modelo de distribuição de dados e seu respectivo classificador como função de uma ploidia definida a priori (Seção 4.2), resta-nos estimar qual a ploidia mais adequada para descrever os dados experimentais e por consequência classificar (ou genotipar) as amostras.

4.3.1 Medidas de distância interna de agrupamentos

Sabendo como genotipar amostra segundo uma ploidia fixa p , precisamos avaliar entre um conjunto finito de possíveis ploidias $\{p_0, p_1, \dots, p_N\}$ qual a mais adequada (verossímil) aos nossos dados.

Pela construção do modelo de genotipagem, conhecemos os padrões esperados na amostra caso ela não sofresse influência de erros experimentais (ver Seção 3.2.1). Como os erros de medida necessariamente ficam em torno desses pontos conhecidos, uma maneira natural de construir um critério de escolha para as ploidias candidatas é a de pensarmos em medidas de distância, como por exemplo:

- Distância Qui-Quadrado: $\chi^2 = \sum_{j=1}^N \frac{(O_j - E_j)^2}{E_j}$;
- Erro Quadrático: $EQ = \sum_{j=1}^N (O_j - E_j)^2$;
- Erro Geométrico Médio $MGE = \left| 1 - \sqrt[N]{\prod_{j=1}^N \left| \frac{1+O_j}{1+E_j} \right|} \right|^2$.

em que,

- O_j : valor observado do j -ésimo elemento da amostra;

- E_j :valor esperado do j-ésimo elemento da amostra.

Aplicando os classificadores descritos na Seção 4.2 para diferentes ploidias, e fazendo uso de medidas de distância como as descritas acima, obtemos uma lista de ploidias candidatas a descrever melhor os dados experimentais. Outras medidas de distância podem ser usadas, vide Apêndice I.1. Como veremos mais adiante na análise de dados simulados Seção 4.6 tais medidas podem ser boas para avaliar a coesão dos grupos em torno de seu ponto central, porém ainda precisamos ter uma idéia da dissimilaridade entre agrupamentos diferentes.

Apesar da simplicidade no critério de escolha da ploidia mais adequada, ainda confrontamos o seguinte problema:

Como garantir que o valor que consideramos ser o esperado para uma determinada observação, é realmente adequado?

Para responder a esta questão, podemos fazer uso de arranjos experimentais com medidas repetidas de um mesmo indivíduo.

4.3.2 Medidas repetidas

Medidas repetidas são comuns em experimentos com cruzamentos controlados. Por exemplo, em um experimento F1, os dados dos pais são medidos várias vezes a fim de verificar a calibração dos instrumentos, pois é esperado que dados do mesmo indivíduo obtidos várias vezes, sob as mesmas condições, possuam algum tipo de estabilidade.

Um critério para avaliar prováveis ploidias, é o de considerar apenas ploidias que comportam medidas coerentes (com classificações iguais) no mesmo genótipo, penalizando ploidias que classificam ambigualmente tais medidas. A Figura 4.6 ilustra este tipo de situação. Dois indivíduos, cada um com 12 medidas repetidas. As regiões de classificação genotípica (dependentes da ploidia) aparecem em branco e cinza claro entre os raios da figura. Um dos indivíduos (marcado com triângulos) possui 11 medidas distribuídas ao longo do eixo P_B e uma única medida no eixo P_A , o que pode ser um indício de erro experimental. O outro indivíduo (marcado com quadrados) fica classificado ambigualmente se consideramos a ploidia 9, porém, na ploidia 10 todas suas medidas são classificadas coerentemente, permanecendo na mesma região de genotipagem (em cinza).

Figura 4.6: Regiões de classificação (genotipagem) com dados reais de cana-de-açúcar.

Na Tabela 4.2, temos para diferentes ploidias (de 2 a 19), os genótipos dos pais (colunas G_1 e G_2) e o número de observações incoerentes (colunas EG_1 e EG_2). Conforme mostrado na Figura (4.6), na ploidia

9, os dois pais possuem cada um, uma medida incoerente, já na ploidia 10, apenas um dos pais possui uma medida incoerente.

p	G_1	G_2	EG_1	EG_2	p	G_1	G_2	EG_1	EG_2
2	1	0	0	1	11	3	0	0	1
3	1	0	0	1	12	4	0	5	1
4	1	0	0	1	13	4	0	0	1
5	1	0	4	1	14	4	0	0	1
6	2	0	0	1	15	4	0	4	1
7	2	0	0	1	16	5	0	2	1
8	2	0	0	1	17	5	0	0	1
9	3	0	1	1	18	5	0	3	1
10	3	0	0	1	19	6	0	5	1

Tabela 4.2: Genótipos e número de medidas incoerentes para diferentes ploidias

4.3.3 Medidas de distância entre de agrupamentos

Na Subseção 4.3.1 vimos como mensurar a coesão ou variabilidade interna dos agrupamentos. Porém, pouca variabilidade ou coesão nos agrupamentos não garante um critério adequado para a estimação da ploidia. Como podemos ver no exemplo descrito pela Figura 4.7, temos um conjunto de dados de ploidia 5. Fazendo variar a ploidia, mudamos os pontos de classificação indicados pelas retas verticais tracejadas. Os valores da densidade sob estes pontos variam conforme a ploidia, e como podemos ver, ploidias com pontos de classificação ruins levam a valores de densidade distantes do eixo x. Para ploidia 5 todos os pontos ficam em cima da abcissa (eixo x).

Figura 4.7: Densidades nos pontos de classificação para diferentes ploidias candidatas

Portanto uma medida natural de distância entre grupos que separe bem as ploidias candidatas é a distância quadrática da densidade estimada nos pontos de fronteira e a abcissa, isto é,

$$D_{sep}(X, p) = \frac{1}{p} \sum_{k=0}^{p-1} f(x_k)^2 \quad (4.10)$$

em que:

- p : ploidia;

- f : densidade dos dados projetados no espaço combinação convexa $[0;1]$ pela transformação U , definida pela Equação 4.3;
- x_k : ponto de fronteira entre agrupamentos diferentes, $x_k = \frac{0,5+k}{p}$ para $k = 0, \dots, p-1$.

A medida D_{sep} assim definida, representa a soma das densidades nos pontos de fronteira. Quanto maior o tamanho da amostra, melhor será a estimação da densidade f e por consequência, melhor a avaliação de D_{sep} . Quanto mais próximo de 0 estiver o valor de D_{sep} é maior o índice de separação dos agrupamentos.

A formulação da Equação 4.10 é referente ao caso de genotipagem de poliplóides bialélicos. Para o caso multialélico, em vez de pontos de fronteira, temos superfícies de fronteira, como vimos na Figura 4.5, sendo necessário o uso de alguma medida resumo a ser aplicada na densidade obtida a partir da amostra para medir a separação entre os agrupamentos.

4.3.4 Estimadores de ploidia

Como nas técnicas de agrupamento, temos uma medida de coesão ou variabilidade interna dos grupos definidas na Subseção 4.3.1, e também temos uma medida de dissimilaridade definida pela Equação 4.10. Assim, podemos definir:

$$D(X, p) = d(X, p) + D_{sep}(X, p) \quad (4.11)$$

Assim, o estimador $\hat{p}$ pode ser definido por:

$$\hat{p} = \arg \min_{p \in \{2, \dots, P\}} D(X, p)$$

- $d(X, p)$, é uma distância de variabilidade interna dos agrupamentos para ploidia p ;
- X , é a matriz de dados da amostra;

Na Seção 4.6 avaliamos em diferentes condições as seguintes medidas:

1. L_1 : distância $EQ + D_{sep}$;
2. L_2 : distância $Qui + D_{sep}$;
3. L_3 : distância $MGE + D_{sep}$;

Obtemos então um critério simples para estimar a ploidia que melhor descreve os dados. De fácil aplicação e que pode ser estendido para dimensões maiores no contexto de genotipagem de multialélicos poliplóides.

4.4 Genotipagem

Uma vez definida a ploidia de um experimento, a genotipagem das amostra é feita automaticamente pelos classificadores definidos em 4.7 e 4.9. Ao utilizar medidas repetidas para estimar a ploidia de um conjunto de dados, podemos estimar também a probabilidade de erros ou acertos na genotipagem destes indivíduos com medidas repetidas, porém, como estimar a probabilidade de genotipagem correta no restante da amostra onde cada indivíduo foi observado somente uma vez?

4.4.1 Estimativa de genotipagem correta

Vamos admitir que o classificador permita apenas erros entre genótipos adjacentes ao verdadeiro, isto é, se um indivíduo possui genótipo $G \in \{0, 1, \dots, p\}$, o classificador 4.7 poderá indicar:

- para $G \in \{1, \dots, p-1\}$, $h_{2,p} = \begin{cases} G-1 \\ G \\ G+1 \end{cases}$,

- para $G = 0$, $h_{2,p} = \begin{cases} 0 \\ 1 \end{cases}$,
- para $G = p$, $h_{2,p} = \begin{cases} p-1 \\ p \end{cases}$.

Assim, podemos construir o modelo de probabilidades condicionais descrito na Tabela 4.3 com as restrições $\sum_{j=0}^p \pi_{i|j} = 1$.

		Genótipo verdadeiro						
		0	1	2	...	...	$p-1$	p
Genótipo Observado	0	$\pi_{0 0}$	$1 - \pi_{0 0}$	-		...	-	-
	1	$\pi_{1 0}$	$\pi_{1 1}$	$\pi_{1 2}$		...	-	-
	2	-	$\pi_{2 1}$	$\pi_{2 2}$			-	-
	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\ddots$	$\ddots$	$\vdots$
	$p-1$	-	-	-	-	$\pi_{p-1 p-2}$	$\pi_{p-1 p-1}$	$\pi_{p-1 p}$
	p	-	-	-	-	-	$1 - \pi_{p p}$	$\pi_{p p}$

Tabela 4.3: Probabilidades condicionais de genotipagem correta com hipótese de erros nas vizinhanças do verdadeiro genótipo.

Descrevendo os eventos:

- $\{L = i\}$: o evento “classificar a observação no genótipo i ”, $i \in \{0, \dots, p\}$ e
- $\{V = j\}$: o evento “a observação pertence ao genótipo j ”, $j \in \{0, \dots, p\}$

As probabilidades $\pi_{i|j} = P(L = i|V = j)$ representam a classificação do elemento amostral condicionada à sua verdadeira classe.

Simplificando o modelo, podemos adicionar a hipótese de simetria,

$$\pi_{i-1|i} = P(L = i-1|V = i) = P(L = i+1|V = i) = \pi_{i+1|i} \quad i \in \{1, \dots, p-1\}$$

e reescrever,

$$\pi_{i|j} = P(L = i|V = j) = \begin{cases} \pi_i & i = j \\ 1 - \pi_i & (i, j) \in \{(0, 1), (p, p-1)\} \\ \frac{1-\pi_i}{2} & i \in \{1, p-1\}, j = i \pm 1 \\ 0 & cc \end{cases} . \quad (4.12)$$

Chegamos então ao modelo descrito pela Tabela 4.4.

As probabilidades de classificação correta π_i podem ser estimadas usando a informação contida em dados de medidas repetidas, porém, nem sempre teremos medidas repetidas suficientes provenientes de diferentes genótipos para estimar todos os elementos da matriz $(\pi_{i|j})$. Para contornar esta situação, consideramos que as probabilidades de classificação correta $\pi_{j|j} = \pi_j$ que aparecem na diagonal da Tabela 4.4 são todas iguais, isto é, $\pi = \pi_i = \pi_j$ com $i \neq j$.

		Genótipo verdadeiro						
		0	1	2	...	...	$p-1$	p
Genótipo Observado	0	π_0	$1 - \pi_0$	-	-	...	-	-
	1	$\frac{1-\pi_1}{2}$	π_1	$\frac{1-\pi_1}{2}$	-	...	-	-
	2		$\frac{1-\pi_2}{2}$	π_2			-	-
	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
	$p-1$	-	-	-	-	$\frac{1-\pi_{p-1}}{2}$	π_{p-1}	$\frac{1-\pi_{p-1}}{2}$
	p	-	-	-	-	-	$1 - \pi_p$	π_p

Tabela 4.4: Probabilidades condicionais de genotipagem correta com hipótese de simetria.

No contexto de medidas repetidas, é esperado que para um mesmo indivíduo, suas observações sejam coerentes, sofrendo pouca variabilidade. Para estimar π podemos usar o número de classificações coerentes (ou seja os genótipos mais observados) das medidas repetidas, ou seja:

$$\pi = \frac{\# \text{ classificações coerentes}}{\# \text{ total de medidas repetidas}} \quad (4.13)$$

Sabendo que podemos escrever a probabilidade de classificação usando o Teorema da Probabilidade Total

$$P(L = i) = \sum_{j=0}^p P(V = j) P(L = i | V = j). \quad (4.14)$$

Como $P(L = i)$ pode ser estimada usando todos os N elementos da amostra (incluindo as medidas não repetidas),

$$P(L = i) = \frac{1}{N} \sum_{l=1}^N \mathbf{1}_{\{G_l=i\}}, \quad (4.15)$$

a Equação 4.14 torna-se um sistema linear tridiagonal. Resolvendo o sistema, encontramos $P(V = j)$ e podemos aplicar o Teorema de Bayes para encontrar a probabilidade de classificação correta da amostra,

$$P(V = i | L = i) = \frac{P(V = i, L = i)}{P(L = i)} \quad (4.16)$$

$$= \frac{P(V = i) P(L = i | V = i)}{P(L = i)} \quad (4.17)$$

$$= \frac{P(V = i) \pi_i}{P(L = i)} \quad (4.18)$$

$$= \pi \frac{P(V = i)}{P(L = i)}. \quad (4.19)$$

A Equação 4.19 nos dá uma medida de probabilidade de genotipagem correta (ou erro de genotipagem) associada à classificação dos indivíduos da amostra em um determinado genótipo i . Note que tal estimativa só foi possível na presença de três hipóteses a respeito da matriz (π_{ij}) :

1. O classificador erra apenas na vizinhança do verdadeiro genótipo;
2. Existe uma simetria nas probabilidades de classificação incorreta e,
3. A probabilidade de classificação correta em cada agrupamento é a mesma, em todos os agrupamentos.

Para termos uma idéia se esta estimativa é razoável, e sob quais condições ela é razoável, mais adiante na Seção 4.6, confrontamos as estimativas avaliadas em simulações com o erro empírico observado.

4.5 Desenhos experimentais e o Equilíbrio de Hardy-Weinberg

As proporções genotípicas observadas nas amostras dependem de basicamente dois fatores:

- **Desenho experimental:** Dependendo do desenho experimental adotado é possível extrair informações adicionais que tornam-se restrições a um modelo de proporções genotípicas das amostras,
- **Condições de cruzamento dos indivíduos observados:** sob hipótese de independência dos cruzamentos e ausência de forças evolutivas (mutações, seleção natural, migrações) podemos adotar a hipótese de que as frequências genotípicas observadas dependem apenas das frequências alélicas populacionais pelo modelo de *Equilíbrio de Hardy-Weinberg* conforme descrito no Apêndice C.

Conforme pode ser visto nos Apêndices B e C, podemos avaliar se as amostras observadas em experimentos de cruzamentos controlados estão em Equilíbrio de Hardy-Weinberg por meio do FBST (Full Bayesian Significance Test).

Nas subseções seguintes veremos alguns exemplos de comportamento das proporções genotípicas amostrais nos casos de Amostra Aleatória Simples (AAS) e cruzamentos F1.

4.5.1 Amostra Aleatória Simples (AAS)

Na Amostra Aleatória Simples (AAS), sob as condições de Equilíbrio de Hardy-Weinberg, as proporções amostrais dos genótipos observados dependem unicamente das proporções alélicas populacionais conforme descrito em C.1.2 e C.1.3. As Tabelas 4.5 e 4.6 ilustram o comportamento das proporções genotípicas em locus dialélicos diplóides e dialélicos tetraplóides, cada par de proporções alélicas populacionais gera um conjunto de proporções genotípicas diferente.

		Genótipos (diplóide)		
f_A	f_B	AA	AB	BB
0,05	0,95	0,0025	0,0950	0,9025
0,30	0,70	0,0900	0,4200	0,4900
0,50	0,50	0,2500	0,5000	0,2500
0,70	0,30	0,4900	0,4200	0,0900
0,95	0,05	0,9025	0,0950	0,0025

Tabela 4.5: Proporções esperadas de genótipos em função da proporção populacional de alelos

		Genótipos (tetraplóide)				
f_A	f_B	$AAAA$	$AAAB$	$AABB$	$ABBB$	$BBBB$
0,05	0,95	0,00000625	0,00047500	0,01353750	0,17147500	0,81450625
0,30	0,70	0,00810000	0,07560000	0,26460000	0,41160000	0,24010000
0,50	0,50	0,06250000	0,25000000	0,37500000	0,25000000	0,06250000
0,70	0,30	0,24010000	0,41160000	0,26460000	0,07560000	0,00810000
0,95	0,05	0,81450625	0,17147500	0,01353750	0,00047500	0,00000625

Tabela 4.6: Proporções esperadas de genótipos em função da proporção populacional de alelos

No caso de um alelo frequente em 5% de uma população, sob as condições de Equilíbrio de Hardy-Weinberg, a proporção esperada de heterozigotos numa população tetraplóide é de somente 625 em cada 10 milhões de indivíduos.

4.5.2 Experimentos F1

Uma vez genotipados os progenitores em um delineamento F1, os agrupamentos genotípicos, assim como suas proporções esperadas podem ser determinadas enumerando os possíveis cruzamentos, de acordo com as leis de segregação de Mendel, como feito no código fonte do Apêndice A que considera os .

Na Tabelas 4.7 e 4.8 ilustramos o resultado para os casos diplóide e triplóide respectivamente, nas duas primeiras linhas, cada entrada da tabela representa a contagem de alelos A nos progenitores, em seguida vem a tabela de proporções, em que cada coluna representa as proporções genotípicas esperadas em possíveis cruzamentos.

Diplóide						
pai	0	0	0	1	1	2
mãe	0	1	2	1	2	2
Alelos	proporções					
0	1,00	0,50	-	0,25	-	-
1	-	0,50	1,00	0,50	0,50	-
2	-	-	-	0,25	0,50	1,00

Tabela 4.7: Proporções genotípicas esperadas no caso diplóide.

Triplóide										
pai	0	0	0	0	1	1	1	2	2	3
mãe	0	1	2	3	1	2	3	2	3	3
Alelos	proporções									
0	1,000	0,500	0,250	-	0,125	0,125	-	0,125	-	-
1	-	0,500	0,500	0,500	0,375	0,375	0,250	0,375	0,250	-
2	-	-	0,250	0,500	0,375	0,375	0,500	0,375	0,500	-
3	-	-	-	-	0,125	0,125	0,250	0,125	0,250	1,000

Tabela 4.8: Proporções genotípicas esperadas no caso triplóide.

Como podemos observar, uma vez genotipados os pais, a genotipagem dos filhos pode ficar restrita, por exemplo: se o pai é homozigoto BBB com 0 alelos A, e a mãe é AAB com 2 alelos B, conforme a Tabela 4.8, os genótipos dos filhos só podem ter 0, 1 ou 2 alelos A.

4.6 Análise de Dados Simulados

A partir das simulações obtidas conforme descrito na Seção 3.5, avaliamos o número de acertos na estimação de ploidia com diferentes distâncias:

1. EQ : distância Erro Quadrático;
2. Qui : distância Qui-quadrado;
3. MGE : distância Erro Geométrico Médio;
4. D_{sep} : distância D_{sep} ;
5. L_1 : distância $EQ + D_{sep}$;
6. L_2 : distância $Qui + D_{sep}$;

7. L_3 : distância $MGE + D_{sep}$;

Uma vez estimada corretamente, avaliamos as medidas resumo da proporção de acertos de genotipagem, isto é, se a genotipagem indicada pelo classificador corresponde ao genótipo original da simulação para as medidas L_1 , L_2 e L_3 .

Nas tabelas da subseção a seguir denotamos por

- G_1 : o genótipo do progenitor 1 (pai), dado pelo número de alelos A;
- G_2 : o genótipo do progenitor 2 (mãe), dado pelo número de alelos A;
- σ : variabilidade ou desvio dos dados projetados no espaço combinação convexa $[0;p]$ em torno de seu genótipo (centróide) esperado.

4.6.1 Cruzamentos F1

simulação	ploidia	G_1	G_2	σ	EQ	Qui	MGE	D_{sep}	L_1	L_2	L_3
1	10	10	7	1.00	1	1	13	29	61	55	40
2	10	10	7	0.75	0	0	19	25	69	67	44
3	10	10	7	0.50	0	0	28	85	324	310	271
4	10	10	7	0.25	56	59	83	0	391	397	488
5	10	10	5	1.00	0	0	25	19	45	37	29
6	10	10	5	0.75	0	0	17	8	30	22	14
7	10	10	5	0.50	0	0	20	125	355	288	235
8	10	10	5	0.25	53	53	79	0	499	499	469
9	10	10	3	1.00	0	0	17	13	44	34	23
10	10	10	3	0.75	0	0	18	6	41	26	21
11	10	10	3	0.50	0	0	33	125	368	332	291
12	10	10	3	0.25	44	47	121	1	489	489	485
13	10	8	3	1.00	0	0	18	4	37	188	34
14	10	8	3	0.75	0	0	17	3	62	48	55
15	10	8	3	0.50	0	0	21	44	368	357	383
16	10	8	3	0.25	53	55	87	0	500	500	500

Tabela 4.9: Número de acertos na estimação de ploidia em simulação de experimentos F1

medida	simulação	G_1	G_2	σ	Mín.	1o.Q	Mediana	Média	3o.Q	Máx.
L_1	1	10	7	1.00	0.44	0.51	0.52	0.53	0.55	0.62
L_2					0.44	0.50	0.52	0.53	0.55	0.62
L_3					0.46	0.50	0.52	0.53	0.55	0.62
L_1	2	10	7	0.75	0.61	0.67	0.69	0.69	0.71	0.77
L_2					0.61	0.67	0.69	0.69	0.71	0.77
L_3					0.62	0.68	0.70	0.70	0.72	0.77
L_1	3	10	7	0.50	0.82	0.93	0.96	0.94	0.97	0.99
L_2					0.82	0.93	0.95	0.94	0.97	0.99
L_3					0.84	0.94	0.96	0.95	0.97	0.99
L_1	4	10	7	0.25	0.92	1.00	1.00	1.00	1.00	1.00
L_2					0.92	1.00	1.00	1.00	1.00	1.00
L_3					0.84	1.00	1.00	0.98	1.00	1.00

Tabela 4.10: Proporção de acertos na simulação de experimentos F1

medida	simulação	G_1	G_2	σ	Mín.	1o.Q	Mediana	Média	3o.Q	Máx.
L_1	5	10	5	1.00	0.41	0.48	0.51	0.50	0.53	0.57
L_2					0.41	0.48	0.51	0.50	0.53	0.57
L_3					0.42	0.49	0.51	0.51	0.53	0.58
L_1	6	10	5	0.75	0.62	0.67	0.69	0.69	0.72	0.76
L_2					0.62	0.67	0.69	0.69	0.72	0.76
L_3					0.63	0.68	0.70	0.69	0.71	0.73
L_1	7	10	5	0.50	0.83	0.94	0.96	0.96	0.97	1.00
L_2					0.91	0.94	0.96	0.96	0.97	1.00
L_3					0.91	0.94	0.96	0.96	0.97	0.99
L_1	8	10	5	0.25	0.98	1.00	1.00	1.00	1.00	1.00
L_2					0.98	1.00	1.00	1.00	1.00	1.00
L_3					0.98	1.00	1.00	1.00	1.00	1.00

Tabela 4.11: Proporção de acertos na simulação de experimentos F1

medida	simulação	G_1	G_2	σ	Mín.	1o.Q	Mediana	Média	3o.Q	Máx.
L_1	5	10	3	1.00	0.42	0.47	0.50	0.50	0.53	0.59
L_2					0.43	0.47	0.50	0.50	0.53	0.59
L_3					0.43	0.48	0.50	0.51	0.54	0.59
L_1	6	10	3	0.75	0.62	0.67	0.70	0.70	0.73	0.77
L_2					0.63	0.68	0.70	0.70	0.73	0.75
L_3					0.63	0.68	0.70	0.70	0.73	0.77
L_1	7	10	3	0.50	0.91	0.95	0.96	0.96	0.97	0.99
L_2					0.91	0.95	0.96	0.96	0.97	0.99
L_3					0.88	0.95	0.96	0.96	0.97	0.99
L_1	8	10	3	0.25	1.00	1.00	1.00	1.00	1.00	1.00
L_2					1.00	1.00	1.00	1.00	1.00	1.00
L_3					1.00	1.00	1.00	1.00	1.00	1.00

Tabela 4.12: Proporção de acertos na simulação de experimentos F1

medida	simulação	G_1	G_2	σ	Mín.	1o.Q	Mediana	Média	3o.Q	Máx.
L_1	13	8	3	1.00	0.43	0.47	0.50	0.50	0.52	0.55
L_2					0.44	0.47	0.50	0.50	0.51	0.55
L_3					0.44	0.47	0.50	0.50	0.52	0.55
L_1	14	8	3	0.75	0.61	0.67	0.69	0.69	0.71	0.76
L_2					0.61	0.67	0.69	0.69	0.71	0.76
L_3					0.63	0.68	0.69	0.69	0.71	0.76
L_1	15	8	3	0.50	0.91	0.94	0.96	0.96	0.97	0.99
L_2					0.91	0.94	0.96	0.96	0.97	0.99
L_3					0.91	0.94	0.96	0.96	0.97	0.99
L_1	16	8	3	0.25	1.00	1.00	1.00	1.00	1.00	1.00
L_2					1.00	1.00	1.00	1.00	1.00	1.00
L_3					1.00	1.00	1.00	1.00	1.00	1.00

Tabela 4.13: Proporção de acertos na simulação de experimentos F1

Observações

Para os 4 cruzamentos simulados no experimento, observamos na Tabela 4.9 que:

1. Individualmente, as medidas de distância internas dos agrupamentos (EQ , Qui e MGE) tem baixo índice de acertos e a distância de separação entre agrupamentos é mais eficiente, porém ainda com um desempenho ruim.
2. As distâncias que combinam os dois tipos de medidas (L_1 , L_2 e L_3) possuem desempenho melhor que as anteriores, e o seu desempenho melhora a medida que os agrupamentos possuem desvio interno (σ) menor, isto é, estão mais separados.
3. A medida com melhor desempenho no experimento de dados simulados de cruzamentos F1 foi a L_3 .

A distribuição da proporção de acertos dos genótipos como podemos observar nas Tabelas 4.10, 4.11, 4.12 e 4.13 muda conforme o desvio interno (σ), quanto menor o desvio, menor a amplitude da distribuição. Para $\sigma=0,25$ observamos que o percentual de acertos foi de 100%.

4.6.2 Amostra aleatória simples

simulação	ploidia	p_A	σ	EQ	Qui	MGE	D_{sep}	L_1	L_2	L_3
1	2	0.30	0.75	0	0	0	30	0	0	0
2	2	0.30	0.50	0	0	0	150	0	0	0
3	2	0.30	0.25	0	0	0	488	19	74	0
4	2	0.30	0.10	0	0	0	500	459	461	0
5	2	0.05	0.75	0	0	0	27	0	0	0
6	2	0.05	0.50	0	0	0	190	0	0	0
7	2	0.05	0.25	0	0	0	487	403	411	0
8	2	0.05	0.10	0	0	0	500	451	438	0
9	6	0.30	0.75	0	0	8	30	26	34	45
10	6	0.30	0.50	0	0	10	36	30	39	36
11	6	0.30	0.25	0	0	6	261	487	488	296
12	6	0.30	0.10	0	0	21	13	500	500	466
13	6	0.05	0.75	0	0	2	0	24	20	10
14	6	0.05	0.50	0	0	0	0	11	9	2
15	6	0.05	0.25	0	0	0	8	75	73	0
16	6	0.05	0.10	0	0	0	0	496	496	0
17	12	0.30	0.75	0	0	20	10	38	33	23
18	12	0.30	0.50	0	0	28	13	50	47	27
19	12	0.30	0.25	0	0	24	151	420	384	271
20	12	0.30	0.10	500	500	73	102	500	500	457
21	12	0.05	0.75	0	0	14	0	0	0	24
22	12	0.05	0.50	0	0	3	0	0	0	31
23	12	0.05	0.25	0	0	0	1	29	28	9
24	12	0.05	0.10	500	500	0	0	496	496	0

Tabela 4.14: Número de acertos na estimação de ploidia em simulação de experimentos AAS

medida	simulação	ploidia	p_A	σ	Mín.	1o.Q	Mediana	Média	3o.Q	Máx.
L_1	1	2	0.30	0.75						
L_2										
L_3										
L_1	2	2	0.30	0.50						
L_2										
L_3										
L_1	3	2	0.30	0.25	0.96	0.97	0.98	0.98	0.99	0.99
L_2					0.95	0.97	0.98	0.98	0.99	0.99
L_3										
L_1	4	2	0.30	0.10	1.00	1.00	1.00	1.00	1.00	1.00
L_2					1.00	1.00	1.00	1.00	1.00	1.00
L_3										
L_1	5	2	0.05	0.75						
L_2										
L_3										
L_1	6	2	0.05	0.50						
L_2										
L_3										
L_1	7	2	0.05	0.25	0.93	0.97	0.98	0.97	0.98	1.00
L_2					0.93	0.97	0.98	0.97	0.98	1.00
L_3										
L_1	8	2	0.05	0.10	1.00	1.00	1.00	1.00	1.00	1.00
L_2					1.00	1.00	1.00	1.00	1.00	1.00
L_3										

Tabela 4.15: Proporção de acertos de genotipagem nas simulações de experimentos AAS

medida	simulação	ploidia	p_A	σ	Mín.	1o.Q	Mediana	Média	3o.Q	Máx.
L_1	9	6	0.30	0.75	0.47	0.49	0.51	0.51	0.53	0.56
L_2					0.41	0.49	0.51	0.50	0.53	0.56
L_3					0.41	0.49	0.50	0.51	0.54	0.59
L_1	10	6	0.30	0.50	0.63	0.68	0.70	0.70	0.72	0.75
L_2					0.63	0.68	0.70	0.70	0.72	0.75
L_3					0.62	0.68	0.70	0.69	0.71	0.74
L_1	11	6	0.30	0.25	0.91	0.94	0.96	0.96	0.97	0.99
L_2					0.91	0.94	0.96	0.96	0.97	0.99
L_3					0.91	0.94	0.96	0.96	0.97	0.99
L_1	12	6	0.30	0.10	1.00	1.00	1.00	1.00	1.00	1.00
L_2					1.00	1.00	1.00	1.00	1.00	1.00
L_3					1.00	1.00	1.00	1.00	1.00	1.00
L_1	13	6	0.05	0.75	0.46	0.50	0.53	0.53	0.55	0.60
L_2					0.46	0.50	0.53	0.53	0.56	0.60
L_3					0.47	0.49	0.49	0.50	0.52	0.56
L_1	14	6	0.05	0.50	0.68	0.70	0.71	0.71	0.73	0.76
L_2					0.68	0.69	0.71	0.70	0.72	0.73
L_3					0.66	0.67	0.67	0.67	0.68	0.68
L_1	15	6	0.05	0.25	0.93	0.96	0.96	0.96	0.97	0.99
L_2					0.93	0.96	0.96	0.96	0.97	0.99
L_3										
L_1	16	6	0.05	0.10	1.00	1.00	1.00	1.00	1.00	1.00
L_2					1.00	1.00	1.00	1.00	1.00	1.00
L_3										

Tabela 4.16: Proporção de acertos de genotipagem nas simulações de experimentos AAS

medida	simulação	ploidia	p_A	σ	Mín.	1o.Q	Mediana	Média	3o.Q	Máx.
L_1	17	12	0.30	0.75	0.41	0.48	0.50	0.50	0.52	0.56
L_2					0.41	0.48	0.51	0.50	0.52	0.56
L_3					0.41	0.48	0.51	0.50	0.52	0.56
L_1	18	12	0.30	0.50	0.62	0.67	0.69	0.69	0.71	0.77
L_2					0.63	0.68	0.69	0.69	0.71	0.77
L_3					0.60	0.67	0.68	0.68	0.71	0.77
L_1	19	12	0.30	0.25	0.90	0.94	0.96	0.96	0.97	0.99
L_2					0.90	0.94	0.96	0.96	0.97	0.99
L_3					0.92	0.94	0.96	0.96	0.97	0.99
L_1	20	12	0.30	0.10	1.00	1.00	1.00	1.00	1.00	1.00
L_2					1.00	1.00	1.00	1.00	1.00	1.00
L_3					1.00	1.00	1.00	1.00	1.00	1.00
L_1	21	12	0.05	0.75						
L_2										
L_3					0.44	0.51	0.53	0.52	0.55	0.58
L_1	22	12	0.05	0.50						
L_2										
L_3					0.62	0.68	0.70	0.69	0.71	0.75
L_1	23	12	0.05	0.25	0.92	0.95	0.96	0.96	0.97	0.98
L_2					0.92	0.95	0.96	0.96	0.97	0.98
L_3					0.91	0.93	0.94	0.94	0.96	0.98
L_1	24	12	0.05	0.10	1.00	1.00	1.00	1.00	1.00	1.00
L_2					1.00	1.00	1.00	1.00	1.00	1.00
L_3										

Tabela 4.17: Proporção de acertos de genotipagem nas simulações de experimentos AAS

Observações

Para os 4 cruzamentos simulados no experimento, observamos na Tabela 4.14 que:

1. Individualmente, as medidas de distância internas dos agrupamentos (EQ , Qui e MGE) tem baixo índice de acertos e a distância de separação entre agrupamentos é mais eficiente e com um desempenho melhor que no experimento de cruzamentos F1.
2. As distâncias que combinam os dois tipos de medidas (L_1 , L_2 e L_3) possuem desempenho melhor que as anteriores, e o seu desempenho melhora a medida que os agrupamentos possuem desvio interno (σ) menor, isto é, estão mais separados.
3. A medida com melhor desempenho no experimento de dados simulados de amostras aleatórias simples foi a L_1 .

A distribuição da proporção de acertos dos genótipos como podemos observar nas Tabelas 4.15, 4.16 e 4.17 muda conforme o desvio interno (σ), quanto menor o desvio, menor a amplitude da distribuição. Para $\sigma = 0,25$ observamos que o percentual de acertos foi de 100%.

4.6.3 Consistência Empírica

Os resultados apresentados nas Subseções 4.6.1 e 4.6.2 evidenciam que o estimador de ploidia $\hat{p}$ proposto é empiricamente consistente:

- Quanto menor a variabilidade dos erros experimentais, ou seja, $Var(\epsilon) \rightarrow 0 \implies \hat{p}$ acerta mais.
- Quanto menor a variabilidade dos erros experimentais, ou seja, $Var(\epsilon) \rightarrow 0 \implies A_G \xrightarrow{d} 1$, onde A_G : número de acertos na genotipagem.

4.7 Fluxo da Análise

Para encerrar o capítulo e revisar o método, podemos descrevê-lo em poucos passos:

1. Transformar os dados (projeção Combinação Convexa)
2. Filtrar possíveis indivíduos que não podem ser classificados
3. Utilizar o desenho experimental como restrição na genotipagem
4. Avaliar a função distância D para diferentes ploidias
5. Testar a hipótese de H-W (FBST) se o desenho experimental requerer o teste.
6. O melhor estimador de p minimiza D satisfazendo H-W
7. Avaliar graficamente a ploidia e os genótipos obtidos.

Capítulo 5

Aplicações

O objetivo deste capítulo é o de apresentar análises com dados de um experimento real. Com o apoio do Prof. Antônio Augusto Franco Garcia, do Departamento de Genética da ESALQ-USP, obtemos dados reais de um experimento com amostras processadas na plataforma MassArray da Sequenom conforme descrito em [9].

5.1 Genotipagem de cana-de-açúcar

5.1.1 A cana-de-açúcar

A cana-de-açúcar cultivada (*S. officinarum*) é um complexo poliplóide e é conhecida como “cana-nobre”, devido aos colmos grossos e suculentos, com alto conteúdo de sacarose e boas características gerais para a industrialização. Já a *S. spontaneum* é uma espécie altamente polimórfica e apresenta alta adaptabilidade a diversas condições ambientais (Matsuoka et al., 1999). Os primeiros trabalhos de melhoramento da cana-de-açúcar envolviam o cruzamento entre estas duas espécies, seguido por retrocruzamentos com *S. officinarum*. No entanto, devido à maior contribuição citológica e gamética desta espécie, houve um aumento da complexidade do genoma da cana-de-açúcar, de modo que os materiais atuais apresentam alto grau de ploidia com freqüente aneuploidia (Souza, 2003).

A Figura 5.1 mostra uma representação esquemática do genoma de um cultivar moderno de cana-de-açúcar deduzido através de experimentos laboratoriais. Observamos a presença de cromossomos provenientes de duas espécies diferentes além de recombinações. Cultivares modernos são altamente poliplóides e aneuplóides com cerca de 120 cromossomos.

Um desafio aos pesquisadores que trabalham no aperfeiçoamento e seleção de linhagens de cana é estimar a ploidia de locos genéticos de interesse e genotipar as amostras adequadamente. Com o apoio do Prof. Antônio Augusto Franco Garcia, do Departamento de Genética da ESALQ-USP, obtemos dados reais de um experimento com amostras processadas na plataforma MassArray da Sequenom.

5.2 Resultados da análise

Os dados foram obtidos com o seguinte desenho experimental:

- 180 observações de cruzamentos F1 independentes
- 12 observações de cada progenitor
- 27 SNPs avaliados

Para cada um dos 27 SNPs, dentro das ploidias variando de 3 a 24, foram calculados:

1. p : ploidia;
2. G_1 : genótipo estimado do progenitor 1 (genótipo mais frequente nas medidas repetidas);

Figura 5.1: Representação do genoma de um cultivar de cana moderno.

3. G_2 : genótipo estimado do progenitor 2 (genótipo mais frequente nas medidas repetidas);
4. EG_1 : número de medidas incoerentes com a classificação do progenitor 1;
5. EG_2 : número de medidas incoerentes com a classificação do progenitor 2;
6. EF : número de medidas incoerentes de filhos classificados no mesmo agrupamento dos pais;
7. EQ : Erro quadrático dos filhos;
8. Qui : Qui-quadrado dos filhos;
9. MGE : distância média geométrica dos filhos;
10. D_{Sep} : distância de separação entre os agrupamentos;
11. $L_1 = EQ + D_{Sep}$: distância L_1 ;
12. $L_2 = Qui + D_{Sep}$: distância L_2 ;
13. $L_3 = MGE + D_{Sep}$: distância L_3 ;
14. $FBST$: Full Bayesian Significance Test para a proporção da progenia classificada;
15. ev_{Cr} : valor crítico do FBST, acima deste valor, rejeitamos a hipótese de Equilíbrio de Hardy-Weinberg;

Todas as medidas citadas foram aplicadas nos dados de áreas e de picos provenientes dos espectros de massa obtidos em cada SNP.

Conforme os gráficos de dispersão para as ploidias estimadas, observamos que a distância $L_1 = EQ_f + D_{sep}$ foi uma medida consistente para estimação de ploidia.

Observações gerais

- Nos gráficos gerados, percebemos uma diferença na variabilidade dos dados dos progenitores em relação aos dados dos filhos, em alguns SNPs, esta diferença é acentuada, o que pode indicar condições experimentais diferenciadas no tratamento das amostras dos filhos e dos pais com medidas repetidas. Apesar disto, o estimador de ploidia mostrou-se adequado, genotipando as amostras de forma coerente, mantendo uma alta densidade em torno dos pontos centróides esperados e baixa densidade próximo aos pontos de fronteira entre os grupos.

- A diferença da ploidia estimada para dados de picos do espectro de massa e de área não invalida a genotipagem das amostras, pois é possível identificar quais grupos tem maior ou menor heterozigocidade e onde a proporção de alelos A ou B aumenta. Para escolha da “melhor” ploidia, podemos considerar as medidas com menor variabilidade interna ou erros quadráticos menores em cada agrupamento.
- As ploidias estimadas variam entre 5 e 14, o que parece ser coerente com os dados de cana-de-açúcar.
- A medida mais visualmente coerente com os gráficos de dispersão foi a distância L_1 .

5.2.1 SNP 4

Figura 5.2: Gráficos descritivos dos dos picos e áreas do espectro de massa: histogramas, distribuições acumuladas empíricas. Os dados dos pais estão em destaque.

Figura 5.3: Gráficos de dispersão e genotipagem para as diferentes medidas aplicadas.

p	G_1	G_2	EG_1	EG_2	EF	EQ	Qui	MGE	D_{sep}	L_1	L_2	L_3	$FBST$	ev_{Cr}
2	2	2	1	3	1	1.7416	0.8708	0.9989	0.0000	1.7416	0.8708	0.9989	0.3290	0.7969
3	3	3	1	3	1	1.5177	0.7645	0.9984	0.1143	1.6320	0.8788	1.1127	0.5342	0.8189
4	4	4	1	3	45	1.7416	0.8708	0.9989	27.8549	29.5965	28.7257	28.8538	0.8000	0.8312
5	5	5	1	7	2	0.7006	0.3854	17.2278	15.8642	16.5648	16.2496	33.0920	0.0000	0.8393
6	6	5	1	4	2	0.3424	0.1858	4.2064	0.2058	0.5481	0.3915	4.4121	0.0006	0.8451
7	7	6	1	3	2	0.1300	0.0713	0.6133	0.0450	0.1750	0.1163	0.6583	0.7500	0.8496
8	8	7	1	3	3	0.1870	0.0990	0.4114	0.0000	0.1870	0.0990	0.4114	0.1171	0.8531
9	9	8	1	3	3	0.1354	0.0737	0.6752	0.0449	0.1803	0.1186	0.7201	0.8000	0.8560
10	10	9	1	3	7	0.2544	0.1334	0.8286	0.1543	0.4087	0.2877	0.9829	0.1180	0.8584
11	11	10	1	3	4	0.2127	0.1144	0.7157	4.8061	5.0188	4.9205	5.5218	0.4998	0.8605
12	12	11	1	3	48	0.3654	0.1903	0.9240	9.4679	9.8333	9.6582	10.3919	0.1389	0.8622
13	13	12	1	3	5	0.2095	0.1121	0.3159	11.6143	11.8238	11.7264	11.9302	0.4991	0.8638
14	14	13	1	4	88	0.4760	0.2466	0.9524	10.7228	11.1988	10.9694	11.6752	0.6000	0.8652
15	15	14	1	7	7	0.1806	0.0962	0.0676	5.6767	5.8574	5.7730	5.7443	0.8600	0.8665
16	16	14	1	6	8	0.1773	0.0942	0.3116	1.8654	2.0427	1.9596	2.1770	0.5924	0.8676
17	17	15	1	4	6	0.1373	0.0745	0.3620	1.0241	1.1614	1.0986	1.3862	0.5234	0.8686
18	18	16	1	4	21	0.1989	0.1050	0.6293	2.4931	2.6919	2.5981	3.1223	0.7774	0.8695
19	19	17	1	3	7	0.1522	0.0820	0.3866	4.9281	5.0803	5.0101	5.3147	0.8500	0.8704
20	20	18	1	3	51	0.2429	0.1276	0.7954	5.8873	6.1302	6.0149	6.6828	0.9026	0.8712

Tabela 5.1: Análise SNP 4 - picos

p	G_1	G_2	EG_1	EG_2	EF	EQ	Qui	MGE	D_{sep}	L_1	L_2	L_3	$FBST$	$evCr$
2	2	2	1	3	1	3.4802	1.7401	0.9999	0.0000	3.4802	1.7401	0.9999	0.3282	0.7969
3	3	3	1	3	1	2.3106	1.2974	2.8299	23.1481	25.4588	24.4455	25.9781	0.5000	0.8189
4	4	4	1	3	99	3.4802	1.7401	0.9999	0.6944	4.1747	2.4345	1.6944	0.8000	0.8312
5	5	5	1	5	2	0.3981	0.2183	0.5546	0.0000	0.3981	0.2183	0.5546	0.0001	0.8393
6	6	5	1	4	3	0.3729	0.2011	0.7378	0.0000	0.3729	0.2011	0.7378	0.0006	0.8451
7	7	6	1	4	3	0.3474	0.1934	0.8894	0.3599	0.7074	0.5533	1.2493	0.7500	0.8496
8	8	7	1	3	30	0.6334	0.3371	0.9563	6.2838	6.9172	6.6209	7.2401	0.1206	0.8531
9	9	8	1	3	4	0.4000	0.2213	0.1273	8.0480	8.4480	8.2693	8.1753	0.7000	0.8560
10	10	9	1	3	88	0.9432	0.4960	0.9876	5.3395	6.2827	5.8355	6.3271	0.1315	0.8584
11	11	10	1	3	6	0.3242	0.1773	0.2667	1.5520	1.8762	1.7293	1.8187	0.5000	0.8605
12	12	11	1	3	108	1.2190	0.6357	0.9948	0.8673	2.0863	1.5031	1.8622	0.1625	0.8622
13	13	12	1	3	28	0.3848	0.2079	0.8102	2.6605	3.0453	2.8683	3.4706	0.5000	0.8638
14	14	13	1	5	149	1.4515	0.7524	0.9972	5.6658	7.1173	6.4182	6.6630	0.6000	0.8652
15	15	14	1	5	68	0.5254	0.2810	0.9359	4.6502	5.1756	4.9312	5.5861	0.8748	0.8665
16	16	15	1	7	180	1.6454	0.8491	0.9983	3.2841	4.9295	4.1332	4.2825	0.2138	0.8676
17	17	15	1	5	15	0.3109	0.1698	0.5249	3.1364	3.4472	3.3062	3.6612	0.9105	0.8686
18	18	16	1	4	120	0.7718	0.4083	0.9797	3.0496	3.8214	3.4579	4.0293	0.7874	0.8695
19	19	17	1	4	39	0.3640	0.1971	0.7794	4.3294	4.6934	4.5264	5.1087	0.8828	0.8704
20	20	18	1	4	157	0.9312	0.4899	0.9874	4.9730	5.9042	5.4629	5.9604	0.9210	0.8712

Tabela 5.2: Análise SNP 4 - areas

5.2.2 SNP 5

Figura 5.4: Gráficos descritivos dos dos picos e áreas do espectro de massa: histogramas, distribuições acumuladas empíricas. Os dados dos pais estão em destaque.

Figura 5.5: Gráficos de dispersão e genotipagem para as diferentes medidas aplicadas.

p	G_1	G_2	EG_1	EG_2	EF	EQ	Qui	MGE	D_{sep}	L_1	L_2	L_3	$FBST$	$evCr$
2	0	0	1	2	4	4.4048	4.4048	4560734.6771	0.0096	4.4144	4.4144	4560734.6867	0.3300	0.7969
3	0	0	1	3	41	4.4048	4.4048	4560734.6771	33.3333	37.7381	37.7381	4560768.0104	0.7500	0.8189
4	0	1	1	4	6	1.9250	1.5420	0.9972	0.1736	2.0986	1.7156	1.1708	0.6000	0.8312
5	0	1	1	2	8	1.4710	1.2275	0.8630	0.0154	1.4864	1.2429	0.8784	0.0000	0.8393
6	0	1	1	2	10	1.4306	1.2265	0.7066	0.0036	1.4342	1.2301	0.7102	0.0004	0.8451
7	0	1	1	2	13	1.5289	1.3381	11.3540	0.0729	1.6018	1.4110	11.4269	0.7500	0.8496
8	0	1	1	2	20	1.6741	1.4884	55.0264	5.8487	7.5228	7.3370	60.8751	0.1151	0.8531
9	0	1	1	3	51	1.8294	1.6467	183.6358	11.1115	12.9409	12.7582	194.7473	0.8000	0.8560
10	0	1	1	6	80	1.9803	1.8004	483.5484	4.6354	6.6157	6.4358	488.1838	0.1175	0.8584
11	0	2	1	5	18	1.4088	1.1935	0.1010	0.6808	2.0896	1.8743	0.7818	0.5000	0.8605
12	0	2	1	4	21	1.4244	1.2213	0.8379	0.2405	1.6649	1.4618	1.0784	0.2974	0.8622
13	0	2	1	3	26	1.4658	1.2706	3.9047	2.3741	3.8399	3.6447	6.2788	0.5000	0.8638
14	0	2	1	2	34	1.5258	1.3353	11.2991	8.3564	9.8822	9.6916	19.6555	0.6000	0.8652
15	0	2	1	3	61	1.5965	1.4088	26.4789	10.8036	12.4001	12.2124	37.2825	0.8750	0.8665
16	0	2	1	4	78	1.6725	1.4868	54.8325	11.5965	13.2690	13.0333	66.4291	0.5204	0.8676
17	0	2	1	6	102	1.7506	1.5665	103.8198	11.1045	12.8552	12.6710	114.9243	0.5077	0.8686
18	0	3	1	6	33	1.4167	1.2147	1.7653	12.8399	14.2566	14.0546	14.6052	0.9399	0.8695
19	0	3	1	4	41	1.4473	1.2501	3.0303	16.7958	18.2431	18.0459	19.8261	0.9000	0.8704
20	0	3	1	4	48	1.4830	1.2896	5.7633	19.7642	21.2471	21.0538	25.5275	0.9754	0.8712

Tabela 5.3: Análise SNP 5 - picos

p	G_1	G_2	EG_1	EG_2	EF	EQ	Qui	MGE	D_{sep}	L_1	L_2	L_3	$FBST$	$evCr$
2	0	0	1	2	5	5.5733	5.5733	100032616.6890	0.9645	6.5379	6.5379	100032617.6535	0.3292	0.7969
3	0	0	1	4	88	5.5733	5.5733	100032616.6890	3.7037	9.2771	9.2771	100032620.3927	0.7500	0.8189
4	0	1	1	2	6	1.5019	1.2095	0.9378	0.0109	1.5127	1.2203	0.9487	0.6000	0.8312
5	0	1	1	2	8	1.3584	1.1363	1.5046	0.0000	1.3584	1.1363	1.5046	0.0000	0.8393
6	0	1	1	2	11	1.5275	1.3130	36.4326	0.4019	1.9294	1.7149	36.8345	0.0005	0.8451
7	0	1	1	3	26	1.7786	1.5588	236.0952	8.8296	10.6081	10.3884	244.9248	0.7500	0.8496
8	0	1	1	3	72	2.0346	1.8102	969.9446	4.8617	6.8963	6.6719	974.8063	0.1176	0.8531
9	0	1	1	4	98	2.2748	2.0481	2656.0063	1.3240	3.5988	3.3721	2657.3303	0.8000	0.8560
10	0	2	1	6	17	1.3393	1.1176	0.6966	0.4109	1.7502	1.5285	1.1074	0.2609	0.8584
11	0	2	1	4	24	1.4052	1.1900	5.9712	3.2992	4.7043	4.4892	9.2704	0.5000	0.8605
12	0	2	1	3	42	1.5082	1.2936	23.9380	4.8770	6.3852	6.1705	28.8150	0.3539	0.8622
13	0	2	1	3	70	1.6294	1.4128	68.3162	3.8648	5.4943	5.2776	72.1810	0.5000	0.8638
14	0	2	1	3	96	1.7581	1.5389	178.1164	4.1478	5.9059	5.6867	182.2642	0.6000	0.8652
15	0	2	1	4	113	1.8886	1.6668	384.1279	5.7413	7.6299	7.4081	389.8692	0.8750	0.8665
16	0	2	1	6	131	2.0170	1.7932	755.6028	9.9932	12.0101	11.7863	765.5960	0.5807	0.8676
17	0	3	1	6	55	1.4300	1.2159	10.1698	10.9927	12.4227	12.2086	21.1625	0.8764	0.8686
18	0	3	2	5	75	1.5026	1.2883	23.1607	12.0521	13.5548	13.3404	35.2128	0.9506	0.8695
19	0	3	2	5	96	1.5833	1.3677	50.0065	13.3301	14.9134	14.6978	63.3367	0.9000	0.8704
20	0	3	2	3	115	1.6686	1.4512	99.4039	15.6448	17.3134	17.0960	115.0486	0.9848	0.8712

Tabela 5.4: Análise SNP 5 - areas

5.2.3 SNP 13

Figura 5.6: Gráficos descritivos dos dos picos e áreas do espectro de massa: histogramas, distribuições acumuladas empíricas. Os dados dos pais estão em destaque.

Figura 5.7: Gráficos de dispersão e genotipagem para as diferentes medidas aplicadas.

p	G_1	G_2	EG_1	EG_2	EF	EQ	Qui	MGE	D_{sep}	L_1	L_2	L_3	$FBST$	ev_{Cr}
2	2	2	0	0	16	4.6919	2.3459	1.0000	39.9488	44.6407	42.2947	40.9488	0.3289	0.7969
3	3	3	0	1	0	4.4583	2.4012	0.9680	11.8367	16.2950	14.2379	12.8047	0.5000	0.8189
4	3	3	0	2	0	2.6765	1.4607	36.4468	0.3121	2.9886	1.7728	36.7589	0.6000	0.8312
5	4	4	0	2	0	1.7494	0.9212	0.1302	0.0000	1.7494	0.9212	0.1302	0.2682	0.8393
6	5	5	0	2	0	2.0137	1.0894	0.5131	13.3163	15.3300	14.4057	13.8294	0.2339	0.8451
7	6	6	0	2	0	1.9289	1.0330	0.4215	0.7061	2.6350	1.7390	1.1276	0.6250	0.8496
8	7	7	0	2	0	1.6845	0.8860	0.8307	0.1561	1.8405	1.0420	0.9868	0.3289	0.8531
9	8	8	0	3	0	1.6068	0.8430	0.8671	23.1870	24.7938	24.0300	24.0541	0.7000	0.8560
10	9	8	2	6	0	1.7132	0.9311	0.0605	36.0788	37.7920	37.0099	36.1393	0.4483	0.8584
11	9	9	1	4	15	1.3645	0.7524	10.6380	17.8443	19.2088	18.5967	28.4823	0.5592	0.8605
12	10	10	0	2	0	1.2559	0.6805	2.4675	2.2425	3.4984	2.9229	4.7100	0.9133	0.8622
13	11	11	0	2	13	1.0847	0.5879	1.1789	0.4233	1.5080	1.0113	1.6022	0.9246	0.8638
14	12	12	0	2	0	1.2869	0.6995	0.0302	5.9181	7.2049	6.6176	5.9483	0.7995	0.8652
15	13	13	0	3	12	1.2711	0.6941	0.2470	5.4005	6.6716	6.0946	5.6475	0.7500	0.8665
16	14	13	0	7	0	1.3557	0.7306	0.1178	5.6471	7.0027	6.3777	5.7648	1.0000	0.8676
17	15	14	2	6	12	1.3041	0.7066	0.4081	19.6149	20.9189	20.3215	20.0230	0.9991	0.8686
18	15	15	3	4	0	1.4008	0.7576	0.5234	14.4808	15.8816	15.2383	15.0042	1.0000	0.8695
19	16	16	0	3	12	1.2938	0.7016	0.5945	3.1498	4.4436	3.8515	3.7444	0.8000	0.8704
20	17	17	0	2	0	1.3242	0.7102	0.1130	2.1847	3.5089	2.8949	2.2977	1.0000	0.8712

Tabela 5.5: Análise SNP 13 - picos

p	G_1	G_2	EG_1	EG_2	EF	EQ	Qui	MGE	D_{sep}	L_1	L_2	L_3	$FBST$	$evCr$
2	2	2	0	0	43	7.2486	3.6243	1.0000	0.6242	7.8728	4.2485	1.6242	0.3298	0.7969
3	2	2	0	3	0	4.8050	2.7238	411.9363	1.1559	5.9610	3.8797	413.0923	0.5000	0.8189
4	3	3	0	2	0	2.6295	1.3989	0.0986	0.0000	2.6295	1.3989	0.0986	0.6000	0.8312
5	4	4	0	2	0	2.6016	1.3938	0.9338	2.2471	4.8487	3.6410	3.1810	0.2097	0.8393
6	5	5	0	2	0	2.7226	1.4643	0.8227	0.3005	3.0231	1.7649	1.1233	0.2403	0.8451
7	6	6	1	2	0	2.4657	1.3124	0.8198	8.6806	11.1463	9.9930	9.5004	0.6250	0.8496
8	6	6	6	5	0	2.1064	1.1434	1.7930	20.2865	22.3929	21.4299	22.0795	0.4448	0.8531
9	7	7	0	3	20	1.9113	1.0808	11.7010	4.5843	6.4957	5.6651	16.2853	0.7000	0.8560
10	8	8	0	3	0	1.8007	0.9856	0.9268	3.2458	5.0465	4.2314	4.1727	0.5924	0.8584
11	9	9	1	2	14	1.5261	0.8358	0.0028	0.9464	2.4725	1.7822	0.9492	0.6404	0.8605
12	10	10	3	3	0	1.9663	1.0803	0.1868	11.7673	13.7336	12.8476	11.9542	0.9926	0.8622
13	10	10	6	6	13	1.9131	1.0604	1.9594	14.1029	16.0160	15.1633	16.0623	0.8321	0.8638
14	11	11	3	4	0	1.9028	1.0336	0.4177	1.0591	2.9619	2.0927	1.4768	0.7959	0.8652
15	12	12	1	3	12	1.7965	0.9850	0.1396	1.0542	2.8507	2.0392	1.1938	0.8125	0.8665
16	13	13	1	3	0	2.0055	1.0942	0.2985	4.3255	6.3310	5.4198	4.6240	0.9994	0.8676
17	14	14	3	5	12	1.9036	1.0401	0.0536	8.8686	10.7722	9.9087	8.9222	0.9749	0.8686
18	14	15	6	6	0	1.9852	1.0760	0.1346	8.0410	10.0262	9.1169	8.1755	1.0000	0.8695
19	15	15	6	5	12	1.9373	1.0594	0.3069	3.8047	5.7421	4.8641	4.1116	0.8499	0.8704
20	16	16	2	3	0	1.9990	1.0819	0.2859	2.1847	4.1837	3.2666	2.4706	1.0000	0.8712

Tabela 5.6: Análise SNP 13 - areas

5.2.4 SNP 37

Figura 5.8: Gráficos descritivos dos dos picos e áreas do espectro de massa: histogramas, distribuições acumuladas empíricas. Os dados dos pais estão em destaque.

Figura 5.9: Gráficos de dispersão e genotipagem para as diferentes medidas aplicadas.

p	G_1	G_2	EG_1	EG_2	EF	EQ	Qui	MGE	D_{sep}	L_1	L_2	L_3	$FBST$	$evCr$
2	0	0	0	0	0	0.9028	0.9028	17448.9914	0.0000	0.9028	0.9028	17448.9914	0.3299	0.7969
3	0	0	0	1	0	0.9028	0.9028	17448.9914	0.0000	0.9028	0.9028	17448.9914	0.7500	0.8189
4	0	0	0	2	6	0.9028	0.9028	17448.9914	0.4877	1.3905	1.3905	17449.4791	0.8000	0.8312
5	0	0	0	2	35	0.9028	0.9028	17448.9914	126.4005	127.3034	127.3034	17575.3920	0.5000	0.8393
6	0	1	0	3	0	1.2974	1.1710	0.8779	1.3004	2.5978	2.4714	2.1783	0.0000	0.8451
7	0	1	0	2	0	0.9820	0.9080	0.6497	0.0000	0.9820	0.9080	0.6497	0.7500	0.8496
8	0	1	0	1	0	0.7913	0.7467	0.2544	0.0000	0.7913	0.7467	0.2544	0.1121	0.8531
9	0	1	0	2	0	0.6787	0.6498	2.4311	0.0000	0.6787	0.6498	2.4311	0.8000	0.8560
10	0	1	0	1	0	0.6111	0.5910	6.7444	0.0000	0.6111	0.5910	6.7444	0.1011	0.8584
11	0	1	0	2	1	0.5707	0.5556	14.1678	0.0528	0.6234	0.6084	14.2205	0.5000	0.8605
12	0	1	0	2	6	0.5381	0.5259	23.6208	0.1806	0.7187	0.7065	23.8014	0.0906	0.8622
13	0	1	0	2	16	0.5257	0.5153	39.0545	3.4449	3.9706	3.9602	42.4994	0.1272	0.8638
14	0	1	0	2	25	0.5014	0.4919	52.0743	17.6978	18.1992	18.1898	69.7721	0.6000	0.8652
15	0	1	0	2	35	0.4777	0.4687	62.5764	49.1269	49.6046	49.5956	111.7033	0.8750	0.8665
16	0	1	0	5	57	0.4851	0.4760	69.1232	18.1119	18.5970	18.5879	87.2351	0.2094	0.8676
17	0	2	0	4	0	0.6487	0.6141	0.5593	13.2212	13.8699	13.8353	13.7805	0.5133	0.8686
18	0	2	0	4	1	0.6060	0.5765	0.5918	33.0777	33.6837	33.6542	33.6695	0.7109	0.8695
19	0	2	0	4	4	0.5584	0.5325	0.4594	18.6025	19.1608	19.1350	19.0618	0.8500	0.8704
20	0	2	0	4	6	0.5202	0.4979	0.8815	11.8988	12.4190	12.3967	12.7803	0.9066	0.8712

Tabela 5.7: Análise SNP 37 - picos

p	G_1	G_2	EG_1	EG_2	EF	EQ	Qui	MGE	D_{sep}	L_1	L_2	L_3	$FBST$	$evCr$
2	0	0	0	0	0	2.9852	2.9852	496858283.9695	0.0000	2.9852	2.9852	496858283.9695	0.3293	0.7969
3	0	0	0	1	35	2.9852	2.9852	496858283.9695	22.3787	25.3640	25.3640	496858306.3483	0.7500	0.8189
4	0	0	0	2	87	2.9852	2.9852	496858283.9695	2.8089	5.7941	5.7941	496858286.7784	0.8000	0.8312
5	0	1	0	4	0	1.8628	1.7309	1.8932	12.2343	14.0971	13.9652	14.1276	0.0034	0.8393
6	0	1	0	2	0	1.2719	1.1984	0.1484	19.4427	20.7145	20.6411	19.5911	0.1675	0.8451
7	0	1	0	1	2	0.8699	0.8218	1.2151	5.8307	6.7006	6.6525	7.0458	0.7500	0.8496
8	0	1	0	1	19	0.6987	0.6598	6.4113	5.8129	6.5116	6.4727	12.2242	0.2166	0.8531
9	0	1	0	1	35	0.5852	0.5453	16.2887	12.4532	13.0383	12.9984	28.7418	0.8000	0.8560
10	0	1	0	2	55	0.5449	0.5015	49.2191	3.2458	3.7908	3.7474	52.4649	0.1786	0.8584
11	0	1	0	2	72	0.6259	0.5792	186.1446	5.2759	5.9018	5.8551	191.4205	0.7500	0.8605
12	0	1	0	2	91	0.7074	0.6572	520.3751	2.4159	3.1233	3.0731	522.7910	0.1767	0.8622
13	0	1	0	3	103	0.7971	0.7439	1246.6640	2.9230	3.7201	3.6669	1249.5869	0.5054	0.8638
14	0	1	0	3	118	0.8695	0.8136	2834.7533	4.2457	5.1152	5.0593	2838.9989	0.8000	0.8652
15	0	2	0	5	35	0.5677	0.5197	8.9298	8.5908	9.1585	9.1106	17.5206	0.8750	0.8665
16	0	2	0	5	52	0.5695	0.5241	19.2030	5.7836	6.3531	6.3078	24.9866	0.8457	0.8676
17	0	2	0	4	65	0.5820	0.5375	33.0164	13.3957	13.9777	13.9332	46.4122	0.7006	0.8686
18	0	2	0	4	80	0.6020	0.5573	45.3899	8.8099	9.4119	9.3672	54.1998	0.9365	0.8695
19	0	2	0	4	96	0.5967	0.5513	60.2081	5.7888	6.3855	6.3401	65.9969	0.9000	0.8704
20	0	2	0	3	115	0.6161	0.5700	112.9311	2.3720	2.9881	2.9420	115.3031	0.9741	0.8712

Tabela 5.8: Análise SNP 37 - areas

5.2.5 SNP 41

Figura 5.10: Gráficos descritivos dos dos picos e áreas do espectro de massa: histogramas, distribuições acumuladas empíricas. Os dados dos pais estão em destaque.

Figura 5.11: Gráficos de dispersão e genotipagem para as diferentes medidas aplicadas.

p	G_1	G_2	EG_1	EG_2	EF	EQ	Qui	MGE	D_{sep}	L_1	L_2	L_3	$FBST$	$euCr$
2	0	0	0	0	0	1.5752	1.5752	98664.4570	0.0000	1.5752	1.5752	98664.4570	0.3296	0.7969
3	0	0	0	0	1	1.5752	1.5752	98664.4570	1.6461	3.2213	3.2213	98666.1031	0.7500	0.8189
4	0	0	1	0	55	1.5752	1.5752	98664.4570	44.4444	46.0196	46.0196	98708.9014	0.8000	0.8312
5	1	0	0	0	0	0.5251	0.4393	0.9970	0.2469	0.7721	0.6862	1.2439	0.0000	0.8393
6	1	0	0	0	0	0.1669	0.1431	0.9631	0.0000	0.1669	0.1431	0.9631	0.0004	0.8451
7	1	0	0	0	0	0.0416	0.0364	0.7327	0.0000	0.0416	0.0364	0.7327	0.7500	0.8496
8	1	0	0	0	0	0.0190	0.0169	0.2121	0.0000	0.0190	0.0169	0.2121	0.1132	0.8531
9	1	0	0	0	1	0.0438	0.0394	2.9943	0.5487	0.5925	0.5881	3.5430	0.8000	0.8560
10	1	0	0	0	6	0.0902	0.0820	9.4826	3.0864	3.1766	3.1684	12.5690	0.1154	0.8584
11	1	0	0	0	24	0.1459	0.1337	22.2522	11.8439	11.9898	11.9776	34.0960	0.5000	0.8605
12	1	0	1	0	55	0.2044	0.1887	44.3945	14.8148	15.0192	15.0035	59.2093	0.1324	0.8622
13	1	0	4	0	76	0.2625	0.2437	79.2482	6.1953	6.4578	6.4391	85.4436	0.1585	0.8638
14	2	0	1	0	0	0.0396	0.0349	0.6310	2.0804	2.1200	2.1154	2.7114	0.6000	0.8652
15	2	0	0	0	1	0.0219	0.0194	0.3659	0.4115	0.4334	0.4309	0.7774	0.8750	0.8665
16	2	0	0	0	4	0.0190	0.0169	0.2121	0.1736	0.1926	0.1905	0.3857	0.4738	0.8676
17	2	0	0	0	7	0.0275	0.0246	1.2747	1.4135	1.4410	1.4381	2.6881	0.5051	0.8686
18	2	0	0	0	17	0.0438	0.0394	2.9943	7.9671	8.0109	8.0065	10.9615	0.6771	0.8695
19	2	0	0	0	33	0.0653	0.0590	5.6292	16.9641	17.0294	17.0231	22.5933	0.9000	0.8704
20	2	0	1	0	55	0.0902	0.0820	9.4826	8.8889	8.9791	8.9709	18.3714	0.8494	0.8712

Tabela 5.9: Análise SNP 41 - picos

p	G_1	G_2	EG_1	EG_2	EF	EQ	Qui	MGE	D_{sep}	L_1	L_2	L_3	$FBST$	ev_{Cr}
2	0	0	0	0	0	0.9615	0.9615	8803.3228	0.0000	0.9615	0.9615	8803.3228	0.3291	0.7969
3	0	0	0	0	0	0.9615	0.9615	8803.3228	0.0000	0.9615	0.9615	8803.3228	0.7500	0.8189
4	0	0	0	0	3	0.9615	0.9615	8803.3228	0.4823	1.4437	1.4437	8803.8050	0.8000	0.8312
5	0	0	4	0	42	0.9615	0.9615	8803.3228	154.3210	155.2824	155.2824	8957.6437	0.5000	0.8393
6	1	0	1	0	0	0.4383	0.3838	0.9868	7.0016	7.4399	7.3854	7.9884	0.0000	0.8451
7	1	0	0	0	0	0.1968	0.1722	0.9762	0.0000	0.1968	0.1722	0.9762	0.7500	0.8496
8	1	0	0	0	0	0.0781	0.0695	0.8918	0.0000	0.0781	0.0695	0.8918	0.1151	0.8531
9	1	0	0	0	0	0.0281	0.0253	0.6436	0.0000	0.0281	0.0253	0.6436	0.8000	0.8560
10	1	0	0	0	0	0.0148	0.0135	0.0646	0.0000	0.0148	0.0135	0.0646	0.1135	0.8584
11	1	0	0	0	0	0.0215	0.0197	1.0749	0.0000	0.0215	0.0197	1.0749	0.5000	0.8605
12	1	0	0	0	3	0.0392	0.0362	3.0507	0.1608	0.2000	0.1970	3.2115	0.1293	0.8622
13	1	0	0	0	12	0.0628	0.0584	6.1609	2.3742	2.4370	2.4325	8.5350	0.1636	0.8638
14	1	0	1	0	23	0.0894	0.0834	10.7011	10.1231	10.2125	10.2065	20.8242	0.6000	0.8652
15	1	0	4	0	42	0.1170	0.1097	16.9448	51.4403	51.5573	51.5500	68.3851	0.8750	0.8665
16	2	0	3	0	0	0.0486	0.0444	0.1534	29.9072	29.9558	29.9516	30.0606	0.4620	0.8676
17	2	0	2	0	0	0.0353	0.0321	0.4628	10.8194	10.8548	10.8516	11.2823	0.5138	0.8686
18	2	0	1	0	0	0.0241	0.0219	0.4345	2.3339	2.3580	2.3558	2.7683	0.7032	0.8695
19	2	0	0	0	1	0.0175	0.0159	0.3152	1.9462	1.9637	1.9621	2.2614	0.9000	0.8704
20	2	0	0	0	3	0.0148	0.0135	0.0646	0.1929	0.2077	0.2064	0.2575	0.8493	0.8712

Tabela 5.10: Análise SNP 41 - areas

5.2.6 SNP 45

Figura 5.12: Gráficos descritivos dos dos picos e áreas do espectro de massa: histogramas, distribuições acumuladas empíricas. Os dados dos pais estão em destaque.

Figura 5.13: Gráficos de dispersão e genotipagem para as diferentes medidas aplicadas.

p	G_1	G_2	EG_1	EG_2	EF	Qui_f	EQ_f	MGE_f	$FBST$	$evCr$	D_{sep}	$EQ_f + D_{sep}$
2	0	0	0	0	0	1.4518	1.4518	37584.8436	0.3297	0.7969	0.0000	1.4518
3	0	0	0	0	0	1.4518	1.4518	37584.8436	0.7500	0.8189	0.0000	1.4518
4	0	0	0	2	48	1.4518	1.4518	37584.8436	0.8000	0.8312	61.1716	62.6234
5	0	0	0	5	82	1.4518	1.4518	37584.8436	0.5000	0.8393	0.2497	1.7015
6	0	0	0	6	87	1.4518	1.4518	37584.8436	0.1429	0.8451	0.0231	1.4749
7	0	0	0	6	88	1.4518	1.4518	37584.8436	0.8750	0.8496	0.0000	1.4518
8	0	1	0	5	0	0.2133	0.2265	0.3328	0.1123	0.8531	0.0000	0.2265
9	0	1	0	5	0	0.2163	0.2284	2.9276	0.8000	0.8560	0.0000	0.2284
10	0	1	0	5	7	0.2408	0.2541	8.4159	0.1068	0.8584	1.1236	1.3776
11	0	1	0	3	26	0.2758	0.2911	18.3829	0.5000	0.8605	8.2877	8.5787
12	0	1	0	5	48	0.3155	0.3329	34.5403	0.1036	0.8622	20.3905	20.7234
13	0	1	0	5	70	0.3569	0.3760	58.5603	0.1503	0.8638	3.6367	4.0127
14	0	1	0	6	80	0.3980	0.4187	91.9445	0.6000	0.8652	0.5896	1.0083
15	0	1	0	7	82	0.4380	0.4600	135.9384	0.8750	0.8665	0.0832	0.5432
16	0	1	0	7	88	0.4763	0.4993	191.4893	0.2074	0.8676	0.2536	0.7529
17	0	2	0	6	10	0.1927	0.2023	1.7731	0.5051	0.8686	4.8526	5.0548
18	0	2	0	6	22	0.2081	0.2191	3.1343	0.6407	0.8695	6.3405	6.5596
19	0	2	0	7	35	0.2193	0.2309	5.5268	0.9000	0.8704	4.9962	5.2271
20	0	1	0	8	136	0.6103	0.6354	537.9584	0.4229	0.8712	12.2343	12.8698

Tabela 5.11: Análise SNP 45 - picos

p	G_1	G_2	EG_1	EG_2	EF	Qui_f	EQ_f	MGE_f	$FBST$	$evCr$	D_{sep}	$EQ_f + D_{sep}$
2	0	0	0	0	0	1.0329	1.0329	7641.2936	0.3284	0.7969	0.0000	1.0329
3	0	0	0	0	1	1.0329	1.0329	7641.2936	0.7500	0.8189	0.0462	1.0791
4	0	0	0	1	15	1.0329	1.0329	7641.2936	0.8000	0.8312	4.9936	6.0265
5	0	0	0	2	59	1.0329	1.0329	7641.2936	0.5000	0.8393	20.2241	21.2570
6	0	0	0	6	82	1.0329	1.0329	7641.2936	0.1429	0.8451	0.5780	1.6109
7	0	0	0	6	87	1.0329	1.0329	7641.2936	0.8750	0.8496	0.2329	1.2658
8	0	0	0	6	88	1.0329	1.0329	7641.2936	0.1112	0.8531	0.0000	1.0329
9	0	1	0	5	1	0.1687	0.1785	0.2014	0.8000	0.8560	0.0154	0.1939
10	0	1	0	5	1	0.1594	0.1673	0.9145	0.1080	0.8584	0.0000	0.1673
11	0	1	0	5	2	0.1663	0.1741	2.9411	0.5000	0.8605	1.7343	1.9083
12	0	1	0	6	15	0.1822	0.1907	6.2264	0.1044	0.8622	1.6645	1.8553
13	0	1	0	6	25	0.2031	0.2126	11.1103	0.1444	0.8638	7.5148	7.7274
14	0	1	0	5	42	0.2266	0.2371	17.8983	0.6000	0.8652	12.8407	13.0777
15	0	1	0	5	60	0.2511	0.2625	26.8436	0.8750	0.8665	6.7506	7.0131
16	0	2	0	7	1	0.1858	0.1980	0.2839	0.4493	0.8676	3.0479	3.2458
17	0	2	0	7	1	0.1669	0.1766	0.1310	0.5074	0.8686	2.0381	2.2147
18	0	2	0	6	2	0.1570	0.1652	0.0864	0.6533	0.8695	0.4401	0.6053
19	0	2	0	6	4	0.1524	0.1597	0.6131	0.9000	0.8704	2.8397	2.9994
20	0	2	0	7	15	0.1522	0.1592	1.3061	0.8348	0.8712	1.9974	2.1566

Tabela 5.12: Análise SNP 45 - areas

5.2.7 SNP 48

Figura 5.14: Gráficos descritivos dos dos picos e áreas do espectro de massa: histogramas, distribuições acumuladas empíricas. Os dados dos pais estão em destaque.

Figura 5.15: Gráficos de dispersão e genotipagem para as diferentes medidas aplicadas.

p	G_1	G_2	EG_1	EG_2	EF	EQ	Qui	MGE	D_{sep}	L_1	L_2	L_3	$FBST$	$evCr$
2	2	2	0	0	2	2.5996	1.2998	0.9998	0.0000	2.5996	1.2998	0.9998	0.3291	0.7969
3	3	3	5	0	1	2.2427	1.1704	0.9943	3.3783	5.6210	4.5487	4.3726	0.5000	0.8189
4	3	4	0	0	2	1.6169	0.8938	9.5862	20.2192	21.8361	21.1130	29.8054	0.6000	0.8312
5	4	5	0	0	1	0.9020	0.5040	9.1262	0.2525	1.1545	0.7565	9.3787	0.0006	0.8393
6	5	6	0	0	3	0.6084	0.3244	0.2772	0.0000	0.6084	0.3244	0.2772	0.0020	0.8451
7	6	7	0	0	3	0.5205	0.2804	0.5532	0.0147	0.5352	0.2951	0.5679	0.7500	0.8496
8	7	8	0	0	13	0.5853	0.3068	0.8650	0.8901	1.4755	1.1969	1.7552	0.1186	0.8531
9	8	9	5	0	4	0.5359	0.2844	0.8078	1.1261	1.6619	1.4105	1.9339	0.7000	0.8560
10	8	10	1	0	5	0.5651	0.3001	0.6126	9.1213	9.6864	9.4215	9.7339	0.2536	0.8584
11	9	11	0	0	4	0.5337	0.2878	0.5160	5.4786	6.0123	5.7664	5.9946	0.3348	0.8605
12	10	12	0	0	8	0.5820	0.3095	0.4512	6.7755	7.3575	7.0851	7.2267	0.3584	0.8622
13	11	13	0	0	6	0.5284	0.2826	0.4055	3.0300	3.5585	3.3126	3.4355	0.5000	0.8638
14	12	14	0	0	20	0.5521	0.2910	0.6616	2.5259	3.0780	2.8169	3.1874	0.6000	0.8652
15	13	15	5	0	8	0.4908	0.2602	0.5945	0.7692	1.2600	1.0293	1.3636	0.8125	0.8665
16	13	16	3	3	8	0.5070	0.2686	0.5868	3.4746	3.9815	3.7432	4.0614	0.7800	0.8676
17	14	17	1	4	85	0.5292	0.2836	3.7536	3.8640	4.3932	4.1476	7.6175	0.8259	0.8686
18	15	18	0	4	17	0.5144	0.2713	0.6860	1.9541	2.4685	2.2253	2.6401	0.9256	0.8695
19	16	19	0	4	87	0.5163	0.2772	2.9651	8.0657	8.5820	8.3429	11.0308	0.8500	0.8704
20	17	20	1	5	27	0.5416	0.2861	0.6807	4.3831	4.9247	4.6692	5.0639	0.9849	0.8712

Tabela 5.13: Análise SNP 48 - picos

p	G_1	G_2	EG_1	EG_2	EF	EQ	Qui	MGE	D_{sep}	L_1	L_2	L_3	$FBST$	ev_{Cr}
2	2	2	0	0	3	3.6346	1.8173	1.0000	0.1578	3.7924	1.9751	1.1578	0.3284	0.7969
3	2	3	0	0	0	2.6933	1.5081	2.2805	26.9326	29.6260	28.4407	29.2132	0.5000	0.8189
4	3	4	0	0	2	1.3595	0.7591	23.4441	2.8603	4.2197	3.6193	26.3044	0.6000	0.8312
5	4	5	0	0	1	0.8109	0.4501	0.8494	0.0000	0.8109	0.4501	0.8494	0.0006	0.8393
6	5	6	0	0	5	0.7289	0.3863	0.7667	0.0526	0.7815	0.4389	0.8193	0.0024	0.8451
7	6	7	4	0	3	0.7214	0.3883	0.8452	1.3287	2.0501	1.7170	2.1739	0.7374	0.8496
8	6	8	1	1	4	0.7710	0.4145	0.7598	5.2175	5.9885	5.6320	5.9773	0.2585	0.8531
9	7	9	0	1	2	0.7769	0.4254	0.4587	8.9775	9.7544	9.4029	9.4362	0.7000	0.8560
10	8	10	0	2	7	0.7793	0.4181	0.5044	1.5781	2.3574	1.9962	2.0825	0.3183	0.8584
11	9	11	0	2	5	0.7095	0.3824	0.5778	2.0841	2.7936	2.4665	2.6619	0.5000	0.8605
12	9	12	5	4	6	0.6703	0.3594	0.5090	1.6555	2.3258	2.0149	2.1645	0.4435	0.8622
13	10	13	1	5	83	0.7787	0.4242	7.9112	2.8084	3.5871	3.2326	10.7196	0.7843	0.8638
14	11	14	0	5	8	0.6937	0.3706	0.5945	2.8166	3.5103	3.1872	3.4111	0.4706	0.8652
15	12	14	0	6	81	0.7105	0.3905	5.0312	5.5455	6.2560	5.9360	10.5767	0.8125	0.8665
16	13	15	0	6	8	0.6666	0.3582	0.5273	3.4776	4.1443	3.8359	4.0049	0.9540	0.8676
17	13	16	4	5	78	0.6681	0.3685	3.6275	1.9026	2.5707	2.2711	5.5300	0.7356	0.8686
18	14	17	2	5	8	0.6634	0.3577	0.5254	3.2841	3.9475	3.6418	3.8095	0.9948	0.8695
19	15	18	0	5	78	0.6237	0.3416	2.6194	4.0532	4.6768	4.3948	6.6726	0.8500	0.8704
20	16	19	0	4	10	0.6123	0.3297	0.4300	4.6001	5.2124	4.9298	5.0302	0.9993	0.8712

Tabela 5.14: Análise SNP 48 - areas

5.2.8 SNP 50

Figura 5.16: Gráficos descritivos dos dos picos e áreas do espectro de massa: histogramas, distribuições acumuladas empíricas. Os dados dos pais estão em destaque.

Figura 5.17: Gráficos de dispersão e genotipagem para as diferentes medidas aplicadas.

p	G_1	G_2	EG_1	EG_2	EF	EQ	Qui	MGE	D_{sep}	L_1	L_2	L_3	$FBST$	$evCr$
2	2	1	0	0	0	4.7697	2.5043	1.0000	0.6173	5.3869	3.1216	1.6172	0.3334	0.7969
3	3	2	0	0	0	2.2753	1.2425	0.0374	0.1829	2.4582	1.4254	0.2203	0.5000	0.8189
4	4	3	0	0	0	1.1752	0.6329	1.8524	69.4444	70.6197	70.0773	71.2968	0.6000	0.8312
5	5	4	0	1	0	0.5695	0.3159	6.5352	0.2469	0.8164	0.5629	6.7821	0.4991	0.8393
6	6	4	0	0	0	0.3353	0.1850	0.7541	0.2058	0.5410	0.3908	0.9598	0.3179	0.8451
7	7	5	0	0	0	0.2611	0.1447	0.4935	17.6367	17.8978	17.7814	18.1301	0.6250	0.8496
8	8	6	0	0	1	0.1564	0.0880	1.2178	3.8966	4.0530	3.9846	5.1144	0.3440	0.8531
9	9	6	0	3	0	0.0861	0.0477	0.1495	0.0610	0.1470	0.1086	0.2105	0.7000	0.8560
10	10	7	0	0	0	0.0965	0.0524	0.6478	0.6173	0.7138	0.6697	1.2651	0.4422	0.8584
11	11	8	0	0	23	0.3029	0.1610	0.2903	1.8356	2.1385	1.9967	2.1259	0.5017	0.8605
12	12	9	0	3	5	0.1602	0.0875	0.3408	26.0059	26.1662	26.0935	26.3468	0.9225	0.8622
13	13	9	0	2	23	0.2255	0.1200	3.6917	16.7349	16.9604	16.8550	20.4266	0.9687	0.8638
14	14	10	0	0	1	0.0622	0.0340	0.8209	4.6683	4.7305	4.7023	5.4892	0.6318	0.8652
15	15	11	0	1	23	0.1277	0.0675	1.3318	0.5670	0.6947	0.6344	1.8987	0.7521	0.8665
16	16	11	0	4	0	0.0417	0.0230	0.0737	2.1026	2.1444	2.1257	2.1763	0.9988	0.8676
17	17	12	0	1	23	0.1071	0.0568	0.7211	4.4071	4.5142	4.4638	5.1281	0.9909	0.8686
18	18	13	0	0	2	0.0436	0.0239	0.2836	4.6555	4.6990	4.6793	4.9390	0.9999	0.8695
19	19	14	0	3	23	0.0905	0.0477	0.5929	5.0810	5.1715	5.1287	5.6739	0.8000	0.8704
20	20	14	0	3	0	0.0393	0.0213	0.1741	14.9383	14.9776	14.9596	15.1124	1.0000	0.8712

Tabela 5.15: Análise SNP 50 - picos

p	G_1	G_2	EG_1	EG_2	EF	EQ	Qui	MGE	D_{sep}	L_1	L_2	L_3	$FBST$	$evCr$
2	2	1	0	0	0	6.3966	3.7125	0.1235	18.6728	25.0695	22.3853	18.7964	0.3334	0.7969
3	3	2	0	0	0	1.8504	1.0725	293.7344	10.9854	12.8358	12.0579	304.7197	0.5000	0.8189
4	4	3	0	0	3	0.5855	0.3332	1.2471	0.1543	0.7398	0.4875	1.4014	0.6033	0.8312
5	5	3	0	5	0	0.4527	0.2585	0.5887	0.9877	1.4403	1.2462	1.5764	0.4943	0.8393
6	6	4	0	0	0	0.3421	0.1975	0.3651	6.2243	6.5664	6.4218	6.5894	0.3099	0.8451
7	7	5	0	0	0	0.2381	0.1348	0.2933	2.2334	2.4714	2.3682	2.5267	0.6517	0.8496
8	8	6	0	3	18	0.3766	0.2115	0.8678	5.8835	6.2601	6.0950	6.7513	0.4003	0.8531
9	9	6	0	1	0	0.2132	0.1212	0.4180	7.9337	8.1468	8.0548	8.3517	0.7251	0.8560
10	10	7	0	0	7	0.1666	0.0952	0.2854	5.3704	5.5370	5.4655	5.6558	0.6872	0.8584
11	11	8	0	2	23	0.2488	0.1341	1.5189	1.3252	1.5741	1.4594	2.8441	0.5775	0.8605
12	12	8	0	3	3	0.0949	0.0540	0.4374	1.7954	1.8903	1.8494	2.2328	0.9822	0.8622
13	13	9	0	0	23	0.1998	0.1081	1.1520	5.1302	5.3300	5.2383	6.2821	0.9987	0.8638
14	14	10	0	1	13	0.1413	0.0807	0.4440	6.3978	6.5391	6.4785	6.8418	0.7529	0.8652
15	15	10	0	5	23	0.1392	0.0746	1.1459	3.2167	3.3559	3.2914	4.3626	0.8197	0.8665
16	16	11	0	1	8	0.0843	0.0482	0.3067	4.3767	4.4610	4.4250	4.6834	1.0000	0.8676
17	17	12	0	0	25	0.1118	0.0603	0.6625	3.4471	3.5590	3.5074	4.1096	0.9998	0.8686
18	18	12	0	6	4	0.0537	0.0307	0.0657	3.8184	3.8722	3.8491	3.8841	1.0000	0.8695
19	19	13	0	3	21	0.0870	0.0466	0.4024	3.6841	3.7710	3.7307	4.0865	0.8041	0.8704
20	20	14	0	0	15	0.0823	0.0473	0.4277	4.8341	4.9164	4.8815	5.2618	1.0000	0.8712

Tabela 5.16: Análise SNP 50 - areas

5.2.9 SNP 60

Figura 5.18: Gráficos descritivos dos dos picos e áreas do espectro de massa: histogramas, distribuições acumuladas empíricas. Os dados dos pais estão em destaque.

Figura 5.19: Gráficos de dispersão e genotipagem para as diferentes medidas aplicadas.

p	G_1	G_2	EG_1	EG_2	EF	EQ	Qui	MGE	D_{sep}	L_1	L_2	L_3	$FBST$	$evCr$
2	2	2	0	0	0	1.9916	0.9958	0.9997	0.0000	1.9916	0.9916	0.9997	0.3294	0.7969
3	3	3	0	0	0	1.6950	0.8845	0.9926	2.9264	4.6214	3.8109	3.9190	0.5000	0.8189
4	3	4	0	0	0	1.1727	0.6088	0.9423	2.7778	3.9505	3.3866	3.7200	0.6000	0.8312
5	4	5	0	1	0	0.8371	0.4370	0.5586	41.7284	42.5655	42.1653	42.2870	0.0001	0.8393
6	5	5	0	6	0	0.5293	0.2811	2.9681	29.6296	30.1590	29.9107	32.5978	0.0222	0.8451
7	6	6	0	1	0	0.3155	0.1688	4.7853	11.2875	11.6030	11.4563	16.0728	0.7500	0.8496
8	7	7	0	1	0	0.1861	0.0993	2.3227	1.0031	1.1891	1.1024	3.3257	0.3086	0.8531
9	8	8	0	1	0	0.1142	0.0612	1.4608	1.4649	1.5791	1.5261	2.9256	0.7000	0.8560
10	8	9	5	1	0	0.0717	0.0384	0.5651	0.4938	0.5655	0.5322	1.0589	0.2738	0.8584
11	9	10	0	1	40	0.3718	0.1954	5.3093	0.6762	1.0480	0.8716	5.9854	0.7291	0.8605
12	10	11	0	1	0	0.0517	0.0275	0.2717	0.9259	0.9777	0.9535	1.1976	0.3266	0.8622
13	11	12	0	1	40	0.2799	0.1461	1.8412	2.1809	2.4607	2.3270	4.0221	0.5402	0.8638
14	12	13	0	1	0	0.0624	0.0329	0.4692	3.1548	3.2172	3.1877	3.6241	0.5295	0.8652
15	13	14	0	1	39	0.2323	0.1210	1.7741	14.5313	14.7637	14.6523	16.3054	0.8125	0.8665
16	14	15	5	2	3	0.0631	0.0334	0.1657	9.1001	9.1633	9.1335	9.2658	0.9336	0.8676
17	14	15	3	6	39	0.1803	0.0936	2.8128	18.7452	18.9255	18.8388	21.5580	0.9296	0.8686
18	15	16	1	6	0	0.0403	0.0213	0.1158	10.2364	10.2768	10.2577	10.3522	0.9845	0.8695
19	16	17	0	4	39	0.1376	0.0712	1.8838	5.1276	5.2652	5.1988	7.0114	0.8500	0.8704
20	17	18	0	1	0	0.0296	0.0157	0.1085	5.0000	5.0296	5.0157	5.1085	0.9987	0.8712

Tabela 5.17: Análise SNP 60 - picos

p	G_1	G_2	EG_1	EG_2	EF	EQ	Qui	MGE	D_{sep}	L_1	L_2	L_3	$FBST$	$evCr$
2	2	2	0	0	0	2.0051	1.0025	0.9997	0.0000	2.0051	1.0025	0.9997	0.3287	0.7969
3	3	3	1	0	0	1.6704	0.8705	0.9923	2.9264	4.5968	3.7969	3.9187	0.5000	0.8189
4	3	4	0	0	0	1.1284	0.5864	0.9314	2.7778	3.9061	3.3641	3.7092	0.6000	0.8312
5	4	5	0	1	0	0.7992	0.4175	0.5411	29.8765	30.6758	30.2940	30.4176	0.0001	0.8393
6	5	5	0	6	0	0.5119	0.2709	1.6700	12.0942	12.6060	12.3651	13.7642	0.0066	0.8451
7	6	6	0	3	0	0.3202	0.1705	2.8555	10.1141	10.4343	10.2846	12.9696	0.7500	0.8496
8	7	7	0	2	0	0.2015	0.1079	2.7341	4.8225	5.0240	4.9304	7.5566	0.3198	0.8531
9	8	8	1	2	0	0.1253	0.0666	1.0354	1.1804	1.3057	1.2470	2.2157	0.7000	0.8560
10	8	9	1	1	0	0.0806	0.0431	0.7174	2.2222	2.3028	2.2653	2.9396	0.2735	0.8584
11	9	10	0	1	40	0.3807	0.2001	6.6311	0.3794	0.7600	0.5795	7.0105	0.7393	0.8605
12	10	11	0	1	0	0.0527	0.0281	0.1707	1.0288	1.0815	1.0569	1.1995	0.3394	0.8622
13	11	12	0	2	40	0.2848	0.1487	1.9538	1.0295	1.3142	1.1781	2.9832	0.5404	0.8638
14	12	13	0	1	0	0.0607	0.0321	0.4015	5.3864	5.4471	5.4185	5.7879	0.5649	0.8652
15	13	14	1	1	40	0.2320	0.1208	1.8840	10.5441	10.7761	10.6649	12.4281	0.8125	0.8665
16	13	15	4	3	0	0.0543	0.0287	0.0655	10.1562	10.2106	10.1850	10.2217	0.9500	0.8676
17	14	16	0	4	40	0.1784	0.0926	2.7423	11.8956	12.0740	11.9882	14.6379	0.9078	0.8686
18	15	16	0	6	0	0.0388	0.0205	0.0660	4.8172	4.8560	4.8377	4.8832	0.9857	0.8695
19	16	17	0	4	40	0.1386	0.0717	2.1918	4.1776	4.3163	4.2494	6.3695	0.8499	0.8704
20	17	18	0	3	0	0.0278	0.0147	0.0729	5.6790	5.7068	5.6937	5.7519	0.9988	0.8712

Tabela 5.18: Análise SNP 60 - areas

5.2.10 SNP 65

Figura 5.20: Gráficos descritivos dos dos picos e áreas do espectro de massa: histogramas, distribuições acumuladas empíricas. Os dados dos pais estão em destaque.

Figura 5.21: Gráficos de dispersão e genotipagem para as diferentes medidas aplicadas.

p	G_1	G_2	EG_1	EG_2	EF	EQ	Qui	MGE	D_{sep}	L_1	L_2	L_3	$FBST$	ev_{Cr}
2	0	0	0	0	1	1.4652	1.4652	45422.0104	0.0000	1.4652	1.4652	45422.0104	0.3291	0.7969
3	0	0	0	0	11	1.4652	1.4652	45422.0104	1.6461	3.1113	3.1113	45423.6565	0.7500	0.8189
4	0	0	0	6	36	1.4652	1.4652	45422.0104	4.9383	6.4035	6.4035	45426.9487	0.8000	0.8312
5	0	1	0	0	0	0.5022	0.4446	0.8125	3.9506	4.4528	4.3952	4.7631	0.0001	0.8393
6	0	1	0	0	1	0.2473	0.2205	0.7664	0.8230	1.0703	1.0436	1.5894	0.0000	0.8451
7	0	1	0	0	1	0.1415	0.1261	0.5323	0.1764	0.3178	0.3025	0.7087	0.7500	0.8496
8	0	1	0	0	5	0.1129	0.1010	0.2726	0.5015	0.6144	0.6025	0.7742	0.1124	0.8531
9	0	1	0	0	11	0.1258	0.1138	2.8447	0.5487	0.6745	0.6625	3.3934	0.8000	0.8560
10	0	1	0	2	20	0.1608	0.1468	8.4043	0.4938	0.6547	0.6406	8.8981	0.1063	0.8584
11	0	1	0	2	27	0.2059	0.1892	18.6831	2.6055	2.8114	2.7947	21.2886	0.5000	0.8605
12	0	1	0	6	37	0.2546	0.2355	35.5973	1.7032	1.9579	1.9387	37.3005	0.1058	0.8622
13	0	2	0	4	3	0.0703	0.0633	1.5841	6.1166	6.1869	6.1799	7.7007	0.4886	0.8638
14	0	2	0	0	8	0.0723	0.0647	1.0272	4.2364	4.3087	4.3011	5.2636	0.5111	0.8652
15	0	2	0	0	12	0.0761	0.0682	1.2096	1.6461	1.7222	1.7143	2.8557	0.8125	0.8665
16	0	2	0	2	16	0.0838	0.0753	1.8231	2.0303	2.1141	2.1056	3.8534	0.4627	0.8676
17	0	2	0	2	21	0.0967	0.0871	2.9051	0.3732	0.4698	0.4603	3.2782	0.5069	0.8686
18	0	2	0	2	26	0.1136	0.1028	5.0833	3.1237	3.2372	3.2265	8.2070	0.6514	0.8695
19	0	2	0	4	38	0.1343	0.1221	8.0627	2.7303	2.8646	2.8524	10.7930	0.8885	0.8704
20	0	2	0	6	45	0.1564	0.1427	9.7883	1.1728	1.3293	1.3155	10.9611	0.8332	0.8712

Tabela 5.19: Análise SNP 65 - picos

p	G_1	G_2	EG_1	EG_2	EF	EQ	Qui	MGE	D_{sep}	L_1	L_2	L_3	$FBST$	$euCr$
2	0	0	0	0	0	1.3238	1.3238	24252.1422	0.0000	1.3238	1.3238	24252.1422	0.3288	0.7969
3	0	0	0	0	9	1.3238	1.3238	24252.1422	0.4115	1.7354	1.7354	24252.5537	0.7500	0.8189
4	0	0	0	3	32	1.3238	1.3238	24252.1422	7.7160	9.0399	9.0399	24259.8583	0.8000	0.8312
5	0	1	0	3	0	0.5178	0.4719	0.1075	15.8025	16.3203	16.2743	15.9099	0.0013	0.8393
6	0	1	0	0	0	0.2804	0.2504	0.8019	1.4632	1.7436	1.7136	2.2651	0.0000	0.8451
7	0	1	0	0	1	0.1597	0.1434	0.6272	0.8098	0.9696	0.9533	1.4371	0.7500	0.8496
8	0	1	0	0	3	0.1126	0.1007	0.2356	0.0772	0.1897	0.1779	0.3127	0.1123	0.8531
9	0	1	0	0	9	0.1113	0.1008	1.2809	0.1372	0.2485	0.2380	1.4181	0.8000	0.8560
10	0	1	0	1	15	0.1348	0.1231	4.5235	0.6173	0.7521	0.7404	5.1408	0.1037	0.8584
11	0	1	0	2	21	0.1698	0.1559	9.5096	1.4117	1.5815	1.5676	10.9213	0.5000	0.8605
12	0	1	0	3	33	0.2097	0.1939	18.5407	2.6177	2.8275	2.8116	21.1585	0.1038	0.8622
13	0	1	0	4	45	0.2516	0.2338	32.1384	1.6240	1.8755	1.8578	33.7624	0.1478	0.8638
14	0	2	0	5	4	0.0590	0.0532	1.3586	2.4349	2.4939	2.4881	3.7936	0.4001	0.8652
15	0	2	0	3	10	0.0648	0.0583	1.4421	5.3498	5.4145	5.4081	6.7919	0.8125	0.8665
16	0	2	0	2	15	0.0722	0.0650	1.5213	1.9772	2.0495	2.0422	3.4986	0.4382	0.8676
17	0	2	0	2	18	0.0827	0.0747	2.2134	2.1362	2.2189	2.2109	4.3496	0.5095	0.8686
18	0	2	0	1	21	0.0951	0.0861	2.7885	1.9247	2.0198	2.0107	4.7131	0.6442	0.8695
19	0	2	0	2	29	0.1106	0.1005	4.3349	2.9332	3.0438	3.0337	7.2681	0.8503	0.8704
20	0	2	0	3	39	0.1278	0.1165	5.9384	3.3951	3.5228	3.5116	9.3334	0.8304	0.8712

Tabela 5.20: Análise SNP 65 - areas

5.2.11 SNP 77

Figura 5.22: Gráficos descritivos dos dos picos e áreas do espectro de massa: histogramas, distribuições acumuladas empíricas. Os dados dos pais estão em destaque.

Figura 5.23: Gráficos de dispersão e genotipagem para as diferentes medidas aplicadas.

p	G_1	G_2	EG_1	EG_2	EF	EQ	Qui	MGE	D_{sep}	L_1	L_2	L_3	$FBST$	ev_{Cr}
2	2	2	0	0	0	1.8947	0.9474	0.9989	0.0000	1.8947	0.9474	0.9989	0.3302	0.7969
3	3	3	0	0	0	1.7334	0.9229	0.9289	22.3787	24.1121	23.3017	23.3076	0.5000	0.8189
4	3	4	0	0	0	0.6213	0.3520	13.0618	0.0000	0.6213	0.3520	13.0618	0.6000	0.8312
5	4	5	0	0	0	0.1571	0.0858	1.1142	0.0000	0.1571	0.0858	1.1142	0.0000	0.8393
6	5	6	0	0	0	0.0467	0.0242	0.4359	0.0000	0.0467	0.0242	0.4359	0.0001	0.8451
7	6	7	0	0	0	0.0658	0.0344	0.7772	0.0000	0.0658	0.0344	0.7772	0.7500	0.8496
8	7	8	0	0	4	0.1337	0.0704	0.8881	0.3511	0.4848	0.4215	1.2393	0.1115	0.8531
9	8	9	0	0	0	0.1645	0.0879	0.7349	7.4596	7.6241	7.5475	8.1944	0.7000	0.8560
10	8	10	6	0	0	0.1264	0.0678	0.2011	6.1172	6.2436	6.1849	6.3183	0.1738	0.8584
11	9	11	0	0	0	0.0742	0.0393	0.1142	3.9628	4.0370	4.0021	4.0770	0.5000	0.8605
12	10	12	0	0	0	0.0462	0.0239	0.4369	0.0462	0.0924	0.0702	0.4832	0.2255	0.8622
13	11	13	0	0	0	0.0452	0.0234	0.6293	0.2739	0.3191	0.2973	0.9032	0.5000	0.8638
14	12	14	0	0	7	0.0649	0.0339	0.7690	0.8044	0.8693	0.8383	1.5733	0.6000	0.8652
15	13	15	0	0	0	0.0629	0.0330	0.6319	4.4757	4.5386	4.5088	5.1077	0.8125	0.8665
16	14	16	1	1	45	0.1324	0.0697	0.8845	11.2356	11.3680	11.3053	12.1201	0.3138	0.8676
17	14	17	5	1	106	0.2282	0.1182	12.8308	3.5983	3.8265	3.7165	16.4291	0.5542	0.8686
18	15	18	2	2	1	0.0420	0.0217	0.5149	1.1345	1.1765	1.1562	1.6494	0.8200	0.8695
19	16	19	0	3	106	0.1645	0.0848	6.8002	0.4348	0.5993	0.5197	7.2350	0.9000	0.8704
20	17	20	0	3	10	0.0495	0.0257	0.7064	0.5618	0.6113	0.5875	1.2681	0.9182	0.8712

Tabela 5.21: Análise SNP 77 - picos

p	G_1	G_2	EG_1	EG_2	EF	EQ	Qui	MGE	D_{sep}	L_1	L_2	L_3	$FBST$	ev_{Cr}
2	2	2	0	0	0	2.3086	1.1543	0.9996	0.0000	2.3086	1.1543	0.9996	0.3293	0.7969
3	2	3	2	0	0	1.7078	0.9761	7.4688	33.7068	35.4146	34.6829	41.1756	0.5000	0.8189
4	3	4	0	0	0	0.4627	0.2607	5.7213	0.0000	0.4627	0.2607	5.7213	0.6000	0.8312
5	4	5	0	0	0	0.1119	0.0593	0.1158	0.0000	0.1119	0.0593	0.1158	0.0000	0.8393
6	5	6	0	0	0	0.0780	0.0402	0.7641	0.0000	0.0780	0.0402	0.7641	0.0001	0.8451
7	6	7	0	0	0	0.1496	0.0790	0.8913	0.0146	0.1641	0.0936	0.9058	0.7500	0.8496
8	7	8	4	0	17	0.2565	0.1353	0.9501	5.6178	5.8743	5.7531	6.5679	0.1114	0.8531
9	7	9	2	0	0	0.1716	0.0925	0.2310	11.2356	11.4072	11.3281	11.4666	0.7000	0.8560
10	8	10	0	1	0	0.0997	0.0526	0.2502	1.9974	2.0971	2.0501	2.2477	0.1770	0.8584
11	9	11	0	1	0	0.0665	0.0342	0.5721	0.0338	0.1002	0.0680	0.6059	0.5000	0.8605
12	10	12	0	3	3	0.0715	0.0370	0.7431	0.4624	0.5338	0.4993	1.2055	0.2285	0.8622
13	11	12	3	6	105	0.4005	0.2094	14.8224	3.2077	3.6082	3.4171	18.0301	0.6203	0.8638
14	11	13	5	5	0	0.0877	0.0460	0.6198	8.0163	8.1040	8.0624	8.6361	0.5701	0.8652
15	12	14	2	5	103	0.2778	0.1447	16.2848	7.0743	7.3521	7.2190	23.3591	0.8125	0.8665
16	13	15	1	5	0	0.0630	0.0325	0.5359	1.6483	1.7113	1.6807	2.1842	0.7191	0.8676
17	14	16	0	5	99	0.1879	0.0973	7.4934	2.5382	2.7261	2.6355	10.0316	0.5499	0.8686
18	15	17	1	5	0	0.0579	0.0299	0.5507	7.0863	7.1442	7.1162	7.6370	0.9184	0.8695
19	16	18	4	5	92	0.1450	0.0752	4.8159	12.6899	12.8349	12.7650	17.5058	0.8500	0.8704
20	16	19	4	5	0	0.0543	0.0281	0.4105	16.6037	16.6581	16.6318	17.0142	0.9935	0.8712

Tabela 5.22: Análise SNP 77 - areas

5.2.12 SNP 79

Figura 5.24: Gráficos descritivos dos dos picos e áreas do espectro de massa: histogramas, distribuições acumuladas empíricas. Os dados dos pais estão em destaque.

Figura 5.25: Gráficos de dispersão e genotipagem para as diferentes medidas aplicadas.

p	G_1	G_2	EG_1	EG_2	EF	EQ	Qui	MGE	D_{sep}	L_1	L_2	L_3	$FBST$	$evCr$
2	0	0	0	0	2	3.3403	3.3403	945773.1453	0.0888	3.4291	3.4291	945773.2341	0.3292	0.7969
3	0	0	0	0	38	3.3403	3.3403	945773.1453	15.1496	18.4899	18.4899	945788.2949	0.7500	0.8189
4	1	0	0	0	3	2.6651	2.2572	0.9791	1.1848	3.8499	3.4420	2.1639	0.6000	0.8312
5	1	0	0	0	4	2.0464	1.7957	0.8778	0.0000	2.0464	1.7957	0.8778	0.0000	0.8393
6	1	0	0	0	5	1.7787	1.6020	0.4172	0.0296	1.8083	1.6316	0.4467	0.0000	0.8451
7	1	0	0	0	11	1.7059	1.5604	7.5210	0.4079	2.1138	1.9683	7.9288	0.7500	0.8496
8	1	0	0	0	20	1.7160	1.5855	32.5365	3.6151	5.3311	5.2006	36.1516	0.1120	0.8531
9	1	0	0	0	43	1.7622	1.6402	97.8293	5.0552	6.8174	6.6954	102.8845	0.8000	0.8560
10	2	0	4	0	8	1.9701	1.7999	0.2803	3.1759	5.1460	4.9758	3.4562	0.2391	0.8584
11	2	0	0	0	10	1.7781	1.5903	0.2051	1.5013	3.2795	3.0916	1.7064	0.5000	0.8605
12	2	0	0	0	15	1.7302	1.5617	1.3666	0.8246	2.5548	2.3863	2.1913	0.3011	0.8622
13	2	0	0	0	22	1.7079	1.5529	3.5351	1.9281	3.6360	3.4809	5.4632	0.5000	0.8638
14	2	0	0	0	34	1.7049	1.5594	6.9529	5.3269	7.0319	6.8863	12.2798	0.6000	0.8652
15	2	0	0	0	48	1.7032	1.5664	15.5441	7.7180	9.4212	9.2844	23.2621	0.8750	0.8665
16	2	0	0	0	71	1.7146	1.5841	30.5638	10.0705	11.7851	11.6546	40.6343	0.4488	0.8676
17	3	0	4	0	19	1.8175	1.6642	1.3298	12.3199	14.1375	13.9841	13.6497	0.7066	0.8686
18	3	0	0	0	25	1.6976	1.5345	2.0772	14.9510	16.6486	16.4856	17.0282	0.9145	0.8695
19	3	0	0	0	30	1.6879	1.5328	3.5367	17.8552	19.5431	19.3881	21.3919	0.8815	0.8704
20	3	0	0	0	43	1.6866	1.5379	5.7348	20.1433	21.8298	21.6812	25.8781	0.9719	0.8712

Tabela 5.23: Análise SNP 79 - picos

p	G_1	G_2	EG_1	EG_2	EF	EQ	Qui	MGE	D_{sep}	L_1	L_2	L_3	$FBST$	$evCr$
2	0	0	0	0	2	2.9111	2.9111	288203.5776	0.0099	2.9209	2.9209	288203.5875	0.3290	0.7969
3	0	0	0	0	18	2.9111	2.9111	288203.5776	11.9701	14.8811	14.8811	288215.5477	0.7500	0.8189
4	1	0	0	0	3	3.0453	2.7640	0.9259	4.2916	7.3369	7.0557	5.2175	0.6000	0.8312
5	1	0	0	0	4	1.9935	1.7345	0.9553	0.0158	2.0093	1.7503	0.9711	0.0000	0.8393
6	1	0	0	0	5	1.6964	1.5169	0.5682	0.0037	1.7001	1.5206	0.5718	0.0001	0.8451
7	1	0	0	0	7	1.5740	1.4322	1.5966	0.0665	1.6405	1.4987	1.6631	0.7500	0.8496
8	1	0	0	0	13	1.5470	1.4239	9.2195	1.1934	2.7404	2.6174	10.4130	0.1128	0.8531
9	1	0	0	0	23	1.5643	1.4518	29.1161	3.9905	5.5547	5.4423	33.1066	0.8000	0.8560
10	1	0	4	0	42	1.6023	1.4966	71.2000	5.7561	7.3584	7.2527	76.9561	0.1084	0.8584
11	2	0	3	0	9	1.7561	1.6176	0.4318	4.6919	6.4480	6.3095	5.1238	0.4581	0.8605
12	2	0	0	0	11	1.7443	1.6165	0.7654	1.5369	3.2813	3.1534	2.3023	0.2730	0.8622
13	2	0	0	0	16	1.5596	1.4117	0.9613	1.2475	2.8070	2.6591	2.2087	0.5000	0.8638
14	2	0	0	0	21	1.5267	1.3918	2.5855	3.5560	5.0826	4.9477	6.1414	0.6000	0.8652
15	2	0	0	0	28	1.5392	1.4100	4.3565	7.2446	8.7838	8.6546	11.6011	0.8750	0.8665
16	2	0	1	0	45	1.5465	1.4235	8.6184	10.9751	12.5217	12.3986	19.5935	0.4539	0.8676
17	2	0	5	0	59	1.5513	1.4342	16.0754	14.1398	15.6911	15.5740	30.2152	0.5049	0.8686
18	3	0	4	0	20	1.6443	1.5100	1.1492	16.8740	18.5183	18.3840	18.0232	0.8977	0.8695
19	3	0	0	0	22	1.6271	1.5010	1.5310	18.0251	19.6522	19.5260	19.5561	0.8500	0.8704
20	3	0	0	0	29	1.6282	1.5069	2.3466	19.6792	21.3074	21.1861	22.0257	0.9727	0.8712

Tabela 5.24: Análise SNP 79 - areas

5.2.13 SNP 82

Figura 5.26: Gráficos descritivos dos dos picos e áreas do espectro de massa: histogramas, distribuições acumuladas empíricas. Os dados dos pais estão em destaque.

Figura 5.27: Gráficos de dispersão e genotipagem para as diferentes medidas aplicadas.

p	G_1	G_2	EG_1	EG_2	EF	EQ	Qui	MGE	D_{sep}	L_1	L_2	L_3	FBS^T	$evCr$
2	2	2	0	0	1	1.9527	0.9763	0.9984	0.0000	1.9527	0.9763	0.9984	0.3296	0.7969
3	3	3	0	0	1	1.5838	0.8069	0.9981	0.0114	1.5953	0.8183	1.0096	0.7180	0.8189
4	4	4	0	4	53	1.9527	0.9763	0.9984	15.9445	17.8971	16.9208	16.9429	0.8000	0.8312
5	5	5	0	4	3	0.7097	0.3889	3.2187	3.0247	3.7344	3.4136	6.2434	0.0000	0.8393
6	6	5	0	6	3	0.4661	0.2532	1.1370	1.2435	1.7096	1.4967	2.3805	0.0001	0.8451
7	7	6	0	5	3	0.2477	0.1414	0.0799	0.3422	0.5898	0.4836	0.4221	0.7500	0.8496
8	8	7	0	3	5	0.3975	0.2116	0.6235	0.0399	0.4374	0.2515	0.6634	0.1118	0.8531
9	9	8	0	3	4	0.3161	0.1752	0.7864	0.0090	0.3251	0.1842	0.7953	0.8000	0.8560
10	10	9	0	3	15	0.4960	0.2608	0.8779	1.9772	2.4732	2.2380	2.8551	0.1028	0.8584
11	11	10	0	5	6	0.3876	0.2116	0.5049	6.5065	6.8940	6.7180	7.0114	0.2670	0.8605
12	12	11	0	5	58	0.6207	0.3236	0.9419	5.3149	5.9356	5.6385	6.2567	0.0958	0.8622
13	13	12	0	5	7	0.3695	0.1994	0.3783	3.6416	4.0111	3.8410	4.0199	0.5000	0.8638
14	14	13	0	5	80	0.7385	0.3828	0.9644	1.9907	2.7292	2.3734	2.9551	0.6000	0.8652
15	15	14	0	5	9	0.3779	0.2019	0.5995	1.0128	1.3907	1.2147	1.6123	0.8263	0.8665
16	16	15	0	6	93	0.8423	0.4346	0.9761	0.9614	1.8037	1.3959	1.9374	0.2044	0.8676
17	17	15	0	7	9	0.3371	0.1839	0.6048	2.4512	2.7883	2.6352	3.0561	0.5134	0.8686
18	18	17	0	7	130	0.9325	0.4794	0.9824	4.4985	5.4310	4.9779	5.4809	0.2414	0.8695
19	19	17	0	8	11	0.3445	0.1865	0.4511	3.8976	4.2421	4.0841	4.3487	0.8500	0.8704
20	20	17	0	8	11	0.3491	0.1882	0.5244	3.3869	3.7359	3.5750	3.9113	0.9550	0.8712

Tabela 5.25: Análise SNP 82 - picos

p	G_1	G_2	EG_1	EG_2	EF	EQ	Qui	MGE	D_{sep}	L_1	L_2	L_3	$FBST$	$evCr$
2	2	2	0	0	2	2.5893	1.2946	0.9995	0.0000	2.5893	1.2946	0.9995	0.3288	0.7969
3	3	3	0	1	1	2.0419	1.0921	0.9771	12.4486	14.4905	13.5406	13.4256	0.5000	0.8189
4	4	4	0	4	72	2.5893	1.2946	0.9995	4.9383	7.5275	6.2329	5.9378	0.8000	0.8312
5	5	5	0	5	3	0.6298	0.3483	2.4796	1.8673	2.4971	2.2156	4.3469	0.0000	0.8393
6	6	5	0	5	4	0.4852	0.2638	0.1873	0.0914	0.5767	0.3553	0.2787	0.0000	0.8451
7	7	6	0	5	4	0.3299	0.1887	0.7060	0.0036	0.3335	0.1923	0.7096	0.7500	0.8496
8	8	7	0	3	7	0.5685	0.3026	0.8797	0.1358	0.7043	0.4384	1.0155	0.1123	0.8531
9	9	8	0	4	5	0.4794	0.2657	0.7705	4.1499	4.6294	4.4156	4.9204	0.7000	0.8560
10	10	9	0	6	55	0.7595	0.3993	0.9589	5.2180	5.9774	5.6173	6.1769	0.1030	0.8584
11	11	10	0	6	7	0.4530	0.2477	0.4655	2.9675	3.4205	3.2152	3.4330	0.4983	0.8605
12	12	11	0	5	78	0.9450	0.4927	0.9806	1.6515	2.5965	2.1442	2.6321	0.0961	0.8622
13	13	12	0	6	8	0.4643	0.2506	0.6892	1.0982	1.5625	1.3489	1.7875	0.5000	0.8638
14	14	13	0	6	96	1.1068	0.5736	0.9886	1.1717	2.2785	1.7453	2.1603	0.6000	0.8652
15	15	14	0	6	30	0.5251	0.2807	0.8245	3.1819	3.7070	3.4626	4.0064	0.8750	0.8665
16	16	15	0	7	138	1.2441	0.6419	0.9923	4.3138	5.5579	4.9557	5.3062	0.1994	0.8676
17	17	16	0	7	62	0.6110	0.3241	0.9111	3.0402	3.6512	3.3643	3.9513	0.5040	0.8686
18	18	15	0	8	11	0.4306	0.2341	0.5800	2.1828	2.6134	2.4169	2.7628	0.8931	0.8695
19	19	16	0	8	103	0.5300	0.2883	3.4865	1.6748	2.2048	1.9631	5.1613	0.8500	0.8704
20	20	19	0	8	179	1.4587	0.7479	0.9953	1.7366	3.1953	2.4845	2.7318	0.4058	0.8712

Tabela 5.26: Análise SNP 82 - areas

5.2.14 SNP 88

Figura 5.28: Gráficos descritivos dos dos picos e áreas do espectro de massa: histogramas, distribuições acumuladas empíricas. Os dados dos pais estão em destaque.

Figura 5.29: Gráficos de dispersão e genotipagem para as diferentes medidas aplicadas.

p	G_1	G_2	EG_1	EG_2	EF	EQ	Qui	MGE	D_{sep}	L_1	L_2	L_3	$FBST$	$evCr$
2	0	0	0	0	0	2.5696	2.5696	1585180.7975	0.0000	2.5696	2.5696	1585180.7975	0.3303	0.7969
3	1	0	2	0	0	1.8983	1.5934	0.7870	16.5066	18.4049	18.1000	17.2936	0.5000	0.8189
4	1	0	0	0	0	0.5459	0.4430	0.9857	0.0000	0.5459	0.4430	0.9857	0.6000	0.8312
5	1	0	0	0	0	0.1207	0.1057	0.5753	0.0000	0.1207	0.1057	0.5753	0.0000	0.8393
6	1	0	0	0	0	0.0677	0.0625	3.4004	0.0000	0.0677	0.0625	3.4004	0.0000	0.8451
7	1	0	0	0	2	0.1426	0.1280	20.3184	0.0468	0.1894	0.1748	20.3652	0.7500	0.8496
8	1	0	1	0	21	0.2585	0.2326	79.0326	9.8765	10.1350	10.1092	88.9092	0.1117	0.8531
9	2	0	2	0	0	0.1619	0.1413	0.2019	5.5022	5.6641	5.6435	5.7041	0.7000	0.8560
10	2	0	0	0	0	0.0971	0.0866	0.0784	1.9753	2.0724	2.0619	2.0537	0.2047	0.8584
11	2	0	0	0	0	0.0621	0.0570	0.7574	2.3421	2.4042	2.3990	3.0995	0.5000	0.8605
12	2	0	0	0	2	0.0624	0.0569	2.4611	2.3777	2.4400	2.4346	4.8387	0.2753	0.8622
13	2	0	1	0	13	0.0891	0.0800	6.0582	5.8222	5.9113	5.9022	11.8804	0.5000	0.8638
14	2	0	4	0	36	0.1330	0.1185	12.0942	7.5244	7.6573	7.6428	19.6186	0.6000	0.8652
15	3	0	2	0	0	0.0528	0.0473	0.1036	5.0937	5.1465	5.1411	5.1973	0.8125	0.8665
16	3	0	1	0	1	0.0418	0.0377	0.0937	3.7712	3.8130	3.8089	3.8649	0.8640	0.8676
17	3	0	0	0	3	0.0398	0.0358	0.4051	1.8015	1.8412	1.8372	2.2066	0.8895	0.8686
18	3	0	0	0	11	0.0472	0.0420	1.1334	3.7672	3.8144	3.8092	4.9006	0.9324	0.8695
19	3	0	1	0	24	0.0642	0.0566	2.4259	5.6534	5.7175	5.7100	8.0793	0.8500	0.8704
20	3	0	5	0	41	0.0888	0.0783	5.0360	6.7901	6.8790	6.8684	11.8261	0.9749	0.8712

Tabela 5.27: Análise SNP 88 - picos

p	G_1	G_2	EG_1	EG_2	EF	EQ	Qui	MGE	D_{sep}	L_1	L_2	L_3	$FBST$	$evCr$
2	0	0	0	0	0	2.0478	2.0478	304135.9104	0.0000	2.0478	2.0478	304135.9104	0.3287	0.7969
3	0	0	6	0	24	2.0478	2.0478	304135.9104	5.5327	7.5805	7.5805	304141.4431	0.7500	0.8189
4	1	0	0	0	0	0.7645	0.6410	0.9744	1.2346	1.9990	1.8756	2.2089	0.6000	0.8312
5	1	0	0	0	0	0.2305	0.1945	0.9185	0.0000	0.2305	0.1945	0.9185	0.0000	0.8393
6	1	0	0	0	0	0.0723	0.0640	0.1557	0.0000	0.0723	0.0640	0.1557	0.0001	0.8451
7	1	0	0	0	1	0.0722	0.0649	3.6745	0.0000	0.0722	0.0649	3.6745	0.7500	0.8496
8	1	0	0	0	7	0.1338	0.1206	16.2747	1.0031	1.1369	1.1237	17.2778	0.1121	0.8531
9	1	0	6	0	24	0.2176	0.1970	42.5953	1.8459	2.0636	2.0429	44.4412	0.8000	0.8560
10	2	0	3	0	0	0.1087	0.0949	0.0076	4.4444	4.5531	4.5393	4.4520	0.2165	0.8584
11	2	0	1	0	0	0.0701	0.0625	0.2135	2.0490	2.1191	2.1114	2.2625	0.4999	0.8605
12	2	0	0	0	1	0.0530	0.0477	0.6352	0.4115	0.4645	0.4592	1.0467	0.2623	0.8622
13	2	0	0	0	6	0.0539	0.0485	1.5850	0.5597	0.6136	0.6082	2.1447	0.5000	0.8638
14	2	0	0	0	11	0.0687	0.0615	3.6537	2.8669	2.9355	2.9284	6.5206	0.6000	0.8652
15	2	0	6	0	24	0.0950	0.0851	7.7772	1.3352	1.4302	1.4203	9.1123	0.8750	0.8665
16	3	0	3	0	0	0.0425	0.0383	1.1849	10.7301	10.7726	10.7685	11.9150	0.6995	0.8676
17	3	0	3	0	1	0.0353	0.0318	0.4932	1.3331	1.3683	1.3649	1.8262	0.6653	0.8686
18	3	0	1	0	6	0.0324	0.0295	1.2049	2.1863	2.2188	2.2158	3.3913	0.8868	0.8695
19	3	0	1	0	9	0.0373	0.0336	1.5851	2.3388	2.3761	2.3725	3.9239	0.8500	0.8704
20	3	0	2	0	16	0.0477	0.0428	2.4061	3.8272	3.8749	3.8699	6.2332	0.9650	0.8712

Tabela 5.28: Análise SNP 88 - areas

5.2.15 SNP 114

Figura 5.30: Gráficos descritivos dos dos picos e áreas do espectro de massa: histogramas, distribuições acumuladas empíricas. Os dados dos pais estão em destaque.

Figura 5.31: Gráficos de dispersão e genotipagem para as diferentes medidas aplicadas.

p	G_1	G_2	EG_1	EG_2	EF	EQ	Qui	MGE	D_{sep}	L_1	L_2	L_3	$FBST$	$evCr$
2	0	0	0	0	2	2.9230	2.9230	6395757.7778	0.0100	2.9330	2.9330	6395757.7877	0.3293	0.7969
3	1	0	0	0	2	2.7706	2.6213	8555.5981	11.7303	14.5009	14.3516	8567.3284	0.5000	0.8189
4	1	0	0	0	3	2.0734	1.9245	1.5694	5.1071	7.1805	7.0316	6.6765	0.6000	0.8312
5	1	0	0	0	4	1.5177	1.4502	0.7134	0.0638	1.5816	1.5141	0.7772	0.0000	0.8393
6	1	0	0	0	5	1.2849	1.2579	12.0441	0.0565	1.3414	1.3144	12.1006	0.0000	0.8451
7	1	0	0	0	9	1.2298	1.2130	74.2621	2.0273	3.2571	3.2403	76.2894	0.7500	0.8496
8	1	0	2	1	14	1.2514	1.2336	254.1380	7.1599	8.4112	8.3935	261.2979	0.1127	0.8531
9	2	0	0	3	6	1.2521	1.2285	59.4566	14.0929	15.3450	15.3214	73.5495	0.7000	0.8560
10	2	0	0	6	7	1.2489	1.2240	21.8266	17.9347	19.1835	19.1587	39.7612	0.2448	0.8584
11	2	1	0	5	7	1.2204	1.1952	9.1002	20.2913	21.5117	21.4865	29.3915	0.2505	0.8605
12	2	1	0	3	8	1.1800	1.1574	4.9921	22.2305	23.4105	23.3880	27.2226	0.3744	0.8622
13	2	1	1	3	9	1.1315	1.1122	2.4428	19.1786	20.3101	20.2908	21.6215	0.5015	0.8638
14	3	1	2	2	9	0.9823	0.9122	1.4218	14.2710	15.2533	15.1832	15.6928	0.4352	0.8652
15	3	1	0	1	10	0.9790	0.9111	0.0227	10.9004	11.8794	11.8116	10.9231	0.7500	0.8665
16	3	1	0	1	11	0.9491	0.8865	0.2682	8.4621	9.4112	9.3486	8.7303	0.9663	0.8676
17	3	1	0	1	12	0.9440	0.8847	0.1790	6.1102	7.0542	6.9949	6.2891	0.8923	0.8686
18	3	1	1	0	13	0.9369	0.8813	0.2707	4.6096	5.5465	5.4909	4.8803	0.9930	0.8695
19	3	1	4	0	14	0.9280	0.8737	0.3163	4.0305	4.9584	4.9041	4.3468	0.8500	0.8704
20	4	1	2	0	13	0.9308	0.8777	0.1836	3.9228	4.8537	4.8006	4.1065	0.9987	0.8712

Tabela 5.29: Análise SNP 114 - picos

p	G_1	G_2	EG_1	EG_2	EF	EQ	Qui	MGE	D_{sep}	L_1	L_2	L_3	$FBST$	evC_r
2	0	0	0	0	6	4.6555	4.6555	5318695061.9695	2.5535	7.2090	7.2090	5318695064.5231	0.3288	0.7969
3	1	0	0	0	2	3.3259	3.0187	94.4121	6.8095	10.1353	9.8282	101.2215	0.5000	0.8189
4	1	0	0	0	3	2.1191	2.0103	23.6345	3.3715	5.4906	5.3818	27.0060	0.6000	0.8312
5	1	0	0	0	4	1.7272	1.6715	43.5983	9.2566	10.9838	10.9281	52.8549	0.0041	0.8393
6	1	0	0	0	8	1.2856	1.1514	21.7749	14.7645	16.0502	15.9160	36.5394	0.1392	0.8451
7	1	0	0	0	26	1.2889	1.1557	34.3949	10.4480	11.7369	11.6037	44.8429	0.7500	0.8496
8	1	0	6	2	51	1.3158	1.1801	61.6851	4.4396	5.7553	5.6197	66.1246	0.2172	0.8531
9	2	0	0	5	8	0.9933	0.8870	0.7330	3.3684	4.3617	4.2554	4.1014	0.8000	0.8560
10	2	1	0	5	6	0.9467	0.8483	0.6067	2.2064	3.1532	3.0547	2.8131	0.3683	0.8584
11	2	1	0	3	7	0.9037	0.8180	0.3518	2.1844	3.0881	3.0023	2.5361	0.5001	0.8605
12	2	1	0	3	9	0.9120	0.8292	0.1629	2.8288	3.7409	3.6580	2.9917	0.7560	0.8622
13	2	1	2	1	11	0.9115	0.8350	0.7674	3.0371	3.9486	3.8721	3.8045	0.5809	0.8638
14	3	1	3	1	9	0.9373	0.8611	1.1104	3.7136	4.6509	4.5747	4.8239	0.6735	0.8652
15	3	1	0	1	10	0.9337	0.8633	0.9336	5.2620	6.1957	6.1252	6.1955	0.8125	0.8665
16	3	1	0	0	13	0.8561	0.7586	1.4444	6.3622	7.2183	7.1207	7.8066	0.9975	0.8676
17	3	1	0	0	14	0.8638	0.7707	0.9992	5.9932	6.8570	6.7639	6.9924	0.9872	0.8686
18	3	1	2	0	16	0.8633	0.7738	0.9280	6.0706	6.9339	6.8444	6.9986	0.9999	0.8695
19	4	1	3	0	13	0.8584	0.7730	1.0209	6.2067	7.0651	6.9796	7.2276	0.8000	0.8704
20	4	1	0	0	15	0.8520	0.7702	1.0198	5.6978	6.5499	6.4681	6.7177	1.0000	0.8712

Tabela 5.30: Análise SNP 114 - areas

5.2.16 SNP 117

Figura 5.32: Gráficos descritivos dos dos picos e áreas do espectro de massa: histogramas, distribuições acumuladas empíricas. Os dados dos pais estão em destaque.

Figura 5.33: Gráficos de dispersão e genotipagem para as diferentes medidas aplicadas.

p	G_1	G_2	EG_1	EG_2	EF	EQ	Qui	MGE	D_{sep}	L_1	L_2	L_3	$FBST$	$evCr$
2	0	0	0	2	0	2.4783	2.4783	9512196.6956	0.0000	2.4783	2.4783	9512196.6956	0.3299	0.7969
3	0	1	0	4	0	1.6802	1.4565	21.7006	1.6461	3.3263	3.1026	23.3467	0.5000	0.8189
4	0	1	0	1	0	0.8738	0.8358	168.7033	0.0000	0.8738	0.8558	168.7033	0.6000	0.8312
5	0	1	0	1	0	0.6852	0.6583	80.4106	2.2222	2.9074	2.8805	82.6329	0.0006	0.8393
6	0	1	0	2	0	0.5503	0.5019	0.2099	48.3768	48.9271	48.8787	48.5867	0.0024	0.8451
7	0	1	0	5	8	0.4243	0.3790	0.5895	8.8723	9.2966	9.2513	9.4618	0.7500	0.8496
8	0	2	0	4	0	0.2728	0.2381	0.9470	5.7099	5.9827	5.9480	6.6569	0.3301	0.8531
9	0	2	0	5	0	0.1303	0.1133	0.9488	0.5487	0.6790	0.6620	1.4975	0.7000	0.8560
10	0	2	0	3	0	0.0514	0.0455	0.7659	0.1235	0.1748	0.1689	0.8893	0.2824	0.8584
11	0	2	0	5	1	0.0355	0.0313	0.0350	0.2263	0.2618	0.2576	0.2613	0.3596	0.8605
12	0	2	0	5	9	0.0579	0.0506	2.4240	0.4230	0.4809	0.4736	2.8470	0.3492	0.8622
13	0	2	0	7	18	0.1023	0.0899	8.3765	2.9940	3.0963	3.0839	11.3705	0.4096	0.8638
14	0	3	0	5	0	0.0586	0.0520	1.3777	9.0559	9.1145	9.1079	10.4336	0.6681	0.8652
15	0	3	0	7	1	0.0561	0.0502	0.6754	1.1065	1.1626	1.1567	1.7819	0.7500	0.8665
16	0	3	0	7	4	0.0576	0.0520	1.5186	14.7811	14.8386	14.8331	16.2996	0.9422	0.8676
17	0	3	0	6	10	0.0675	0.0605	1.4520	22.1710	22.2385	22.2314	23.6229	0.9036	0.8686
18	0	4	0	7	0	0.0487	0.0439	0.0267	17.9173	17.9661	17.9612	17.9440	0.9969	0.8695
19	0	4	0	7	0	0.0423	0.0378	0.3540	8.2948	8.3371	8.3326	8.6488	0.7500	0.8704
20	0	4	0	7	1	0.0337	0.0300	0.3952	3.8272	3.8609	3.8572	4.2224	0.9998	0.8712

Tabela 5.31: Análise SNP 117 - picos

p	G_1	G_2	EG_1	EG_2	EF	EQ	Qui	MGE	D_{sep}	L_1	L_2	L_3	$FBST$	$evCr$
2	0	0	0	1	0	1.4086	1.4086	156059.8308	0.0000	1.4086	1.4086	156059.8308	0.3288	0.7969
3	0	1	0	6	0	1.3700	1.3261	20830.6063	2.9264	4.2964	4.2525	20833.5326	0.5000	0.8189
4	0	1	0	2	0	0.8299	0.7332	3.3503	0.3086	1.1386	1.0419	3.6590	0.6000	0.8312
5	0	1	0	1	0	0.4641	0.4401	19.5785	0.0000	0.4641	0.4401	19.5785	0.0035	0.8393
6	0	1	0	1	0	0.3493	0.3451	80.8279	0.0914	0.4407	0.4366	80.9193	0.0025	0.8451
7	0	1	0	2	0	0.3283	0.3218	76.2241	18.6589	18.9871	18.9807	94.8830	0.7500	0.8496
8	0	1	0	6	0	0.3103	0.2960	14.9734	52.8164	53.1267	53.1123	67.7898	0.1115	0.8531
9	0	2	0	6	0	0.2769	0.2565	0.2561	30.0378	30.3146	30.2943	30.2938	0.8000	0.8560
10	0	2	0	4	0	0.2190	0.1995	0.8291	13.0864	13.3054	13.2859	13.9155	0.2685	0.8584
11	0	2	0	4	0	0.1489	0.1361	0.8744	8.0975	8.2464	8.2336	8.9719	0.4281	0.8605
12	0	2	0	4	0	0.0879	0.0800	0.9103	1.2460	1.3339	1.3260	2.1563	0.3301	0.8622
13	0	2	0	5	0	0.0515	0.0471	0.8225	0.1821	0.2336	0.2292	1.0046	0.4065	0.8638
14	0	2	0	6	3	0.0368	0.0335	0.5783	0.0882	0.1249	0.1217	0.6665	0.4214	0.8652
15	0	3	0	6	0	0.0294	0.0265	0.3612	0.5853	0.6147	0.6118	0.9464	0.8125	0.8665
16	0	3	0	5	0	0.0282	0.0253	0.0325	4.3403	4.3684	4.3656	4.3728	0.7526	0.8676
17	0	3	0	7	0	0.0256	0.0228	0.0754	1.8155	1.8411	1.8384	1.8909	0.6953	0.8686
18	0	3	0	6	0	0.0236	0.0213	0.3036	1.9010	1.9246	1.9223	2.2046	0.8975	0.8695
19	0	4	0	7	0	0.0237	0.0216	0.5409	2.4128	2.4365	2.4344	2.9537	0.8500	0.8704
20	0	4	0	7	0	0.0231	0.0212	0.7125	8.0864	8.1095	8.1076	8.7989	0.9845	0.8712

Tabela 5.32: Análise SNP 117 - areas

5.2.17 SNP 122

Figura 5.34: Gráficos descritivos dos dos picos e áreas do espectro de massa: histogramas, distribuições acumuladas empíricas. Os dados dos pais estão em destaque.

Figura 5.35: Gráficos de dispersão e genotipagem para as diferentes medidas aplicadas.

p	G_1	G_2	EG_1	EG_2	EF	EQ	Qui	MGE	D_{sep}	L_1	L_2	L_3	$FBST$	$euCr$
2	0	0	0	0	0	0.9144	0.9144	16109.3248	0.0000	0.9144	0.9144	16109.3248	0.3301	0.7969
3	0	0	0	0	0	0.9144	0.9144	16109.3248	0.0000	0.9144	0.9144	16109.3248	0.7500	0.8189
4	0	0	0	0	6	0.9144	0.9144	16109.3248	0.3086	1.2230	1.2230	16109.6335	0.8000	0.8312
5	0	0	0	0	27	0.9144	0.9144	16109.3248	63.2099	64.1242	64.1242	16172.5347	0.5000	0.8393
6	1	0	4	0	0	0.5064	0.4629	0.5484	24.8971	25.4035	25.3600	25.4456	0.0000	0.8451
7	1	0	0	0	0	0.2923	0.2636	0.9149	7.2886	7.5810	7.5523	8.2035	0.7500	0.8496
8	1	0	0	0	0	0.1533	0.1386	0.8764	0.6173	0.7706	0.7558	1.4937	0.1179	0.8531
9	1	0	0	0	0	0.0819	0.0747	0.6881	0.2049	0.2868	0.2796	0.8930	0.8000	0.8560
10	1	0	0	0	0	0.0504	0.0465	0.2015	0.1235	0.1738	0.1699	0.3250	0.1319	0.8584
11	1	0	0	0	3	0.0430	0.0400	0.8928	0.3348	0.3778	0.3749	1.2276	0.5000	0.8605
12	1	0	0	0	6	0.0499	0.0467	2.9070	0.1029	0.1528	0.1496	3.0099	0.1485	0.8622
13	1	0	0	0	8	0.0651	0.0610	6.2426	2.9945	3.0597	3.0556	9.2372	0.1766	0.8638
14	1	0	0	0	17	0.0850	0.0798	11.3290	4.1464	4.2314	4.2262	15.4755	0.6000	0.8652
15	1	0	0	0	27	0.1073	0.1010	18.5937	21.1065	21.2138	21.2076	39.7003	0.8750	0.8665
16	1	0	1	0	45	0.1306	0.1234	28.4371	9.5100	9.6406	9.6334	37.9471	0.2145	0.8676
17	1	0	4	0	60	0.1542	0.1461	41.2135	6.9418	7.0960	7.0879	48.1553	0.3255	0.8686
18	2	0	4	0	3	0.0416	0.0388	0.7794	8.8113	8.8529	8.8501	9.5907	0.7009	0.8695
19	2	0	4	0	5	0.0376	0.0351	0.7196	6.0249	6.0624	6.0600	6.7445	0.8500	0.8704
20	2	0	1	0	6	0.0357	0.0334	0.8363	6.7901	6.8258	6.8235	7.6265	0.8892	0.8712

Tabela 5.33: Análise SNP 122 - picos

p	G_1	G_2	EG_1	EG_2	EF	EQ	Qui	MGE	D_{sep}	L_1	L_2	L_3	$FBST$	$euCr$
2	0	0	0	0	0	0.9145	0.9145	16113.2342	0.0000	0.9145	0.9145	16113.2342	0.3294	0.7969
3	0	0	0	0	0	0.9145	0.9145	16113.2342	0.0000	0.9145	0.9145	16113.2342	0.7500	0.8189
4	0	0	0	0	6	0.9145	0.9145	16113.2342	0.3086	1.2232	1.2232	16113.5428	0.8000	0.8312
5	0	0	0	0	27	0.9145	0.9145	16113.2342	71.3580	72.2726	72.2726	16184.5922	0.5000	0.8393
6	1	0	4	0	0	0.5057	0.4632	0.4730	28.0064	28.5121	28.4696	28.4794	0.0001	0.8451
7	1	0	0	0	0	0.2923	0.2636	0.9149	6.6551	6.9474	6.9187	7.5700	0.7500	0.8496
8	1	0	0	0	0	0.1534	0.1391	0.8610	0.3472	0.5007	0.4863	1.2082	0.1178	0.8531
9	1	0	0	0	0	0.0823	0.0754	0.6533	0.5487	0.6310	0.6241	1.2020	0.8000	0.8560
10	1	0	0	0	0	0.0509	0.0470	0.2013	0.1235	0.1744	0.1704	0.3248	0.1298	0.8584
11	1	0	0	0	3	0.0435	0.0405	0.8932	0.3348	0.3783	0.3753	1.2281	0.5000	0.8605
12	1	0	0	0	6	0.0504	0.0471	2.9080	0.1029	0.1533	0.1500	3.0108	0.1492	0.8622
13	1	0	0	0	8	0.0656	0.0614	6.2444	2.9945	3.0601	3.0560	9.2389	0.1729	0.8638
14	1	0	0	0	17	0.0854	0.0802	11.3320	4.1464	4.2318	4.2266	15.4784	0.6000	0.8652
15	1	0	0	0	27	0.1077	0.1014	18.5985	23.8226	23.9303	23.9240	42.4211	0.8750	0.8665
16	1	0	1	0	45	0.1310	0.1238	28.4442	7.8897	8.0207	8.0134	36.3339	0.2114	0.8676
17	1	0	4	0	59	0.1546	0.1464	41.2238	6.0877	6.2423	6.2341	47.3114	0.3222	0.8686
18	2	0	4	0	3	0.0417	0.0389	0.8735	9.8478	9.8894	9.8867	10.7213	0.6983	0.8695
19	2	0	4	0	5	0.0380	0.0355	0.7200	7.7845	7.8225	7.8200	8.5045	0.8500	0.8704
20	2	0	1	0	6	0.0360	0.0336	0.7533	5.6173	5.6533	5.6509	6.3706	0.8963	0.8712

Tabela 5.34: Análise SNP 122 - areas

5.2.18 SNP 136

Figura 5.36: Gráficos descritivos dos dos picos e áreas do espectro de massa: histogramas, distribuições acumuladas empíricas. Os dados dos pais estão em destaque.

Figura 5.37: Gráficos de dispersão e genotipagem para as diferentes medidas aplicadas.

p	G_1	G_2	EG_1	EG_2	EF	EQ	Qui	MGE	D_{sep}	L_1	L_2	L_3	$FBST$	$evCr$
2	0	0	0	0	0	0.5018	0.5018	840.0177	0.0000	0.5018	0.5018	840.0177	0.3282	0.7969
3	0	0	0	0	0	0.5018	0.5018	840.0177	0.0000	0.5018	0.5018	840.0177	0.7500	0.8189
4	0	0	0	0	1	0.5018	0.5018	840.0177	0.0000	0.5018	0.5018	840.0177	0.8000	0.8312
5	0	0	0	0	9	0.5018	0.5018	840.0177	6.3123	6.8142	6.8142	846.3301	0.5000	0.8393
6	0	0	0	2	27	0.5018	0.5018	840.0177	3.6530	4.1548	4.1548	843.6707	0.1429	0.8451
7	0	0	0	6	49	0.5018	0.5018	840.0177	22.1070	22.6088	22.6088	862.1247	0.8750	0.8496
8	0	1	0	0	0	0.5719	0.5315	0.4963	3.9452	4.5171	4.4767	4.4415	0.1114	0.8531
9	0	1	0	0	0	0.4933	0.4630	0.5257	2.7708	3.2642	3.2339	3.2966	0.8000	0.8560
10	0	1	0	0	0	0.4138	0.3924	0.2726	0.7890	1.2028	1.1815	1.0617	0.1008	0.8584
11	0	1	0	0	0	0.3672	0.3515	0.2321	0.0000	0.3672	0.3515	0.2321	0.5000	0.8605
12	0	1	0	0	1	0.3311	0.3194	1.0780	0.0000	0.3311	0.3194	1.0780	0.0835	0.8622
13	0	1	0	0	1	0.3072	0.2980	2.2431	0.2299	0.5371	0.5279	2.4729	0.1202	0.8638
14	0	1	0	0	6	0.2917	0.2842	3.7595	0.5176	0.8093	0.8018	4.2771	0.6000	0.8652
15	0	1	0	0	9	0.2794	0.2729	5.2319	2.1041	2.3835	2.3771	7.3360	0.8750	0.8665
16	0	1	0	1	14	0.2735	0.2678	7.3908	0.6241	0.8977	0.8919	8.0150	0.1858	0.8676
17	0	1	0	1	19	0.2705	0.2653	9.9198	3.8561	4.1266	4.1213	13.7758	0.2299	0.8686
18	0	1	0	2	27	0.2696	0.2646	12.8115	1.2177	1.4872	1.4823	14.0292	0.2285	0.8695
19	0	1	0	4	33	0.2701	0.2654	16.0529	6.8206	7.0907	7.0860	22.8735	0.9000	0.8704
20	0	1	0	5	40	0.2717	0.2671	19.6262	7.9890	8.2607	8.2562	27.6152	0.3705	0.8712

Tabela 5.35: Análise SNP 136 - picos

p	G_1	G_2	EG_1	EG_2	EF	EQ	Qui	MGE	D_{sep}	L_1	L_2	L_3	$FBST$	$evCr$
2	0	0	0	0	0	0.3969	0.3969	344.0208	0.0000	0.3969	0.3969	344.0208	0.3291	0.7969
3	0	0	0	0	0	0.3969	0.3969	344.0208	0.0000	0.3969	0.3969	344.0208	0.7500	0.8189
4	0	0	0	0	0	0.3969	0.3969	344.0208	0.0000	0.3969	0.3969	344.0208	0.8000	0.8312
5	0	0	0	0	4	0.3969	0.3969	344.0208	0.0000	0.3969	0.3969	344.0208	0.5000	0.8393
6	0	0	0	0	15	0.3969	0.3969	344.0208	14.6119	15.0088	15.0088	358.6326	0.1429	0.8451
7	0	0	0	1	35	0.3969	0.3969	344.0208	11.1340	11.5309	11.5309	355.1548	0.8750	0.8496
8	0	0	0	1	48	0.3969	0.3969	344.0208	6.1644	6.5613	6.5613	350.1851	0.1112	0.8531
9	0	0	0	4	60	0.3969	0.3969	344.0208	7.8904	8.2873	8.2873	351.9112	0.9000	0.8560
10	0	0	0	6	68	0.3969	0.3969	344.0208	12.6247	13.0215	13.0215	356.6454	0.0910	0.8584
11	0	1	0	3	0	0.3420	0.3250	0.3984	0.2134	0.5554	0.5385	0.6119	0.5000	0.8605
12	0	1	0	1	0	0.3049	0.2919	0.1475	0.0000	0.3049	0.2919	0.1475	0.0816	0.8622
13	0	1	0	0	1	0.2750	0.2650	0.3304	0.0000	0.2750	0.2650	0.3304	0.1193	0.8638
14	0	1	0	0	1	0.2542	0.2464	0.9526	0.0000	0.2542	0.2464	0.9526	0.6000	0.8652
15	0	1	0	0	4	0.2399	0.2334	1.7270	0.0000	0.2399	0.2334	1.7270	0.8750	0.8665
16	0	1	0	0	6	0.2302	0.2247	2.6574	0.4932	0.7234	0.7178	3.1506	0.1894	0.8676
17	0	1	0	0	10	0.2238	0.2189	3.7433	3.8962	4.1200	4.1152	7.6395	0.2392	0.8686
18	0	1	0	0	15	0.2181	0.2137	4.6661	5.2170	5.4351	5.4307	9.8831	0.2227	0.8695
19	0	1	0	0	22	0.2159	0.2118	5.9958	1.5852	1.8011	1.7970	7.5810	0.9000	0.8704
20	0	1	0	0	26	0.2149	0.2111	7.4617	8.0877	8.3026	8.2988	15.5494	0.3653	0.8712

Tabela 5.36: Análise SNP 136 - areas

5.2.19 SNP 151

Figura 5.38: Gráficos descritivos dos dos picos e áreas do espectro de massa: histogramas, distribuições acumuladas empíricas. Os dados dos pais estão em destaque.

Figura 5.39: Gráficos de dispersão e genotipagem para as diferentes medidas aplicadas.

p	G_1	G_2	EG_1	EG_2	EF	EQ	Qui	MGE	D_{sep}	L_1	L_2	L_3	$FBST$	$evCr$
2	2	2	0	0	5	3.1064	1.5532	0.9998	0.6173	3.7237	2.1705	1.6171	0.3295	0.7969
3	3	3	0	0	0	1.6479	0.8997	0.1038	7.7275	9.3754	8.6272	7.8313	0.5000	0.8189
4	3	4	0	0	1	0.5973	0.3375	16.8113	1.9290	2.5263	2.2665	18.7404	0.6000	0.8312
5	4	5	0	0	0	0.1776	0.0987	1.7858	0.0000	0.1776	0.0987	1.7858	0.0000	0.8393
6	5	6	0	0	6	0.2109	0.1150	0.5163	0.2058	0.4166	0.3208	0.7221	0.0000	0.8451
7	6	7	0	0	1	0.1369	0.0763	0.3521	0.7568	0.8936	0.8330	1.1088	0.6250	0.8496
8	7	8	0	0	34	0.4712	0.2513	0.9345	1.1671	1.6383	1.4184	2.1016	0.1125	0.8531
9	8	9	0	0	2	0.1405	0.0771	0.4137	2.5775	2.7180	2.6547	2.9912	0.7000	0.8560
10	8	10	5	0	6	0.1507	0.0826	0.3929	2.6235	2.7741	2.7061	3.0163	0.2372	0.8584
11	9	11	0	0	2	0.0903	0.0503	0.2744	3.1845	3.2748	3.2348	3.4589	0.5001	0.8605
12	10	12	0	0	24	0.2042	0.1113	0.5951	1.3296	1.5338	1.4409	1.9247	0.2852	0.8622
13	11	13	0	0	3	0.0742	0.0414	0.0234	1.1629	1.2372	1.2043	1.1864	0.5029	0.8638
14	12	14	0	0	46	0.3218	0.1733	0.8466	0.7581	1.0799	0.9314	1.6047	0.6000	0.8652
15	13	15	0	0	13	0.1027	0.0568	0.4624	1.8221	1.9248	1.8789	2.2846	0.8125	0.8665
16	14	16	1	0	77	0.4712	0.2513	0.9345	1.5643	2.0355	1.8156	2.4989	0.4089	0.8676
17	14	17	4	0	95	0.3771	0.1970	11.8841	2.3230	2.7001	2.5199	14.2071	0.7677	0.8686
18	15	18	0	0	38	0.1939	0.1057	0.6719	2.9317	3.1256	3.0374	3.6036	0.8957	0.8695
19	16	19	0	0	104	0.3297	0.1726	6.8778	1.6765	2.0061	1.8491	8.5543	0.8500	0.8704
20	17	20	0	0	61	0.2739	0.1481	0.8057	1.1806	1.4545	1.3286	1.9863	0.9643	0.8712

Tabela 5.37: Análise SNP 151 - picos

p	G_1	G_2	EG_1	EG_2	EF	EQ	Qui	MGE	D_{sep}	L_1	L_2	L_3	$FBST$	$evCr$
2	2	2	0	0	8	3.2563	1.6281	0.9998	2.4691	5.7254	4.0973	3.4690	0.3297	0.7969
3	2	3	0	0	0	1.6183	0.8904	0.8760	4.1267	5.7449	5.0170	5.0026	0.5000	0.8189
4	3	4	0	0	1	0.5623	0.3167	11.7734	0.9452	1.5075	1.2619	12.7186	0.6000	0.8312
5	4	5	0	0	0	0.1801	0.1001	1.2832	0.0617	0.2418	0.1618	1.3450	0.0000	0.8393
6	5	6	0	0	9	0.2423	0.1322	0.6036	0.8230	1.0653	0.9552	1.4266	0.0000	0.8451
7	6	7	0	0	1	0.1445	0.0806	0.3768	0.6083	0.7528	0.6888	0.9850	0.6250	0.8496
8	7	8	4	0	39	0.5322	0.2839	0.9464	2.6259	3.1581	2.9097	3.5722	0.1119	0.8531
9	7	9	0	0	1	0.1126	0.0620	0.3032	1.3756	1.4881	1.4375	1.6787	0.7000	0.8560
10	8	10	0	0	9	0.1660	0.0912	0.4456	2.4691	2.6351	2.5603	2.9147	0.2421	0.8584
11	9	11	0	0	2	0.0854	0.0477	0.1573	2.5276	2.6129	2.5753	2.6849	0.5004	0.8605
12	10	12	0	0	26	0.2332	0.1271	0.6826	1.0195	1.2528	1.1466	1.7021	0.2921	0.8622
13	11	13	4	0	4	0.0873	0.0487	0.2316	1.9259	2.0132	1.9746	2.1575	0.5038	0.8638
14	11	14	6	0	10	0.0988	0.0550	0.3210	1.3268	1.4256	1.3818	1.6478	0.5001	0.8652
15	12	15	0	0	93	0.4624	0.2418	17.9016	1.0745	1.5369	1.3163	18.9761	0.8102	0.8665
16	13	16	0	0	21	0.1517	0.0835	0.5899	1.5468	1.6985	1.6303	2.1367	0.7729	0.8676
17	14	17	1	0	96	0.3865	0.2023	11.3894	2.0809	2.4674	2.2832	13.4703	0.7524	0.8686
18	15	18	2	0	42	0.2241	0.1222	0.7145	3.3991	3.6232	3.5213	4.1136	0.9036	0.8695
19	16	19	4	0	108	0.3443	0.1808	6.2563	1.5420	1.8863	1.7228	7.7983	0.8500	0.8704
20	16	20	5	0	21	0.1099	0.0609	0.4680	0.7562	0.8660	0.8171	1.2242	0.9930	0.8712

Tabela 5.38: Análise SNP 151 - areas

5.2.20 SNP 162

Figura 5.40: Gráficos descritivos dos dos picos e áreas do espectro de massa: histogramas, distribuições acumuladas empíricas. Os dados dos pais estão em destaque.

Figura 5.41: Gráficos de dispersão e genotipagem para as diferentes medidas aplicadas.

p	G_1	G_2	EG_1	EG_2	EF	EQ	Qui	MGE	D_{sep}	L_1	L_2	L_3	$FBST$	ev_{Cr}
2	2	2	0	0	2	2.5255	1.2628	0.9996	0.0000	2.5255	1.2628	0.9996	0.3289	0.7969
3	3	3	0	0	2	1.9700	1.0294	0.9995	0.0457	2.0157	1.0751	1.0452	0.7216	0.8189
4	4	4	0	3	63	2.5255	1.2628	0.9996	39.9354	42.4609	41.1981	40.9350	0.8000	0.8312
5	5	4	0	3	3	0.8678	0.5016	2.1583	1.8673	2.7351	2.3688	4.0256	0.0000	0.8393
6	6	5	0	0	5	0.9108	0.4922	0.3990	0.0432	0.9540	0.5354	0.4422	0.0001	0.8451
7	7	6	0	0	5	0.6310	0.3590	0.7628	0.0020	0.6331	0.3611	0.7648	0.7500	0.8496
8	8	7	0	0	7	0.8745	0.4633	0.9029	0.0074	0.8819	0.4706	0.9103	0.1117	0.8531
9	9	8	0	0	7	0.7416	0.4098	0.9449	0.1043	0.8459	0.5141	1.0492	0.8000	0.8560
10	10	9	0	1	12	0.9921	0.5197	0.9669	0.7716	1.7637	1.2913	1.7385	0.1023	0.8584
11	11	10	1	3	9	0.8735	0.4740	0.9185	3.7638	4.6373	4.2378	4.6823	0.4882	0.8605
12	12	11	3	3	70	1.1255	0.5855	0.9798	13.9764	15.1019	14.5619	14.9562	0.1103	0.8622
13	12	11	5	6	91	0.7371	0.4183	12.9473	13.9287	14.6658	14.3470	26.8760	0.5538	0.8638
14	13	12	4	5	11	0.7181	0.3973	0.5119	7.7041	8.4222	8.1014	8.2160	0.6000	0.8652
15	14	13	3	3	76	0.6727	0.3802	2.6915	6.5786	7.2514	6.9589	9.2701	0.8750	0.8665
16	15	14	2	1	13	0.7346	0.4022	0.6297	8.5448	9.2795	8.9470	9.1745	0.7582	0.8676
17	16	15	1	2	57	0.6814	0.3815	0.3935	10.7806	11.4620	11.1621	11.1742	0.8565	0.8686
18	17	16	1	3	15	0.7714	0.4187	0.4309	13.2768	14.0482	13.6955	13.7077	0.9506	0.8695
19	18	17	1	3	38	0.7040	0.3904	0.3867	17.8396	18.5436	18.2300	18.2263	0.8500	0.8704
20	19	18	1	3	17	0.7840	0.4226	0.1519	23.1817	23.9656	23.6043	23.3336	0.9966	0.8712

Tabela 5.39: Análise SNP 162 - picos

p	G_1	G_2	EG_1	EG_2	EF	EQ	Qui	MGE	D_{sep}	L_1	L_2	L_3	FBS^T	$evCr$
2	2	2	0	0	2	3.1691	1.5845	0.9999	0.0000	3.1691	1.5845	0.9999	0.3288	0.7969
3	3	3	0	2	2	2.5709	1.3556	0.9995	5.7870	8.3580	7.1427	6.7865	0.5000	0.8189
4	4	3	0	6	3	1.5787	0.8864	4.4111	7.3351	8.9138	8.2214	11.7462	0.6000	0.8312
5	5	4	0	2	3	0.7768	0.4443	0.4289	0.0000	0.7768	0.4443	0.4289	0.0000	0.8393
6	6	5	0	0	5	0.9375	0.5021	0.8706	0.0804	1.0178	0.5824	0.9510	0.0000	0.8451
7	7	6	0	0	5	0.7334	0.4107	0.9423	0.3422	1.0756	0.7528	1.2845	0.7500	0.8496
8	8	7	0	0	8	1.0320	0.5448	0.9706	0.6602	1.6923	1.2050	1.6308	0.1135	0.8531
9	9	8	1	2	7	0.9203	0.5042	0.9459	4.8106	5.7309	5.3148	5.7565	0.7431	0.8560
10	10	9	3	3	53	1.2046	0.6314	0.9721	17.2242	18.4287	17.8555	18.1962	0.1608	0.8584
11	11	10	4	4	9	0.8738	0.4754	0.3873	24.1600	25.0338	24.6354	24.5472	0.2649	0.8605
12	11	10	6	6	9	0.6926	0.3889	0.2345	16.9528	17.6454	17.3417	17.1873	0.5142	0.8622
13	12	11	5	4	37	0.5992	0.3458	1.8508	14.8965	15.4958	15.2424	16.7474	0.5455	0.8638
14	13	12	5	3	11	0.6980	0.3875	0.4134	12.3406	13.0386	12.7281	12.7540	0.6000	0.8652
15	14	13	5	4	26	0.6298	0.3586	0.0396	11.5746	12.2044	11.9333	11.6143	0.8750	0.8665
16	15	14	5	3	13	0.7393	0.4060	0.2530	12.6044	13.3437	13.0104	12.8574	0.9418	0.8676
17	16	15	4	3	21	0.6556	0.3689	0.3532	15.3639	16.0194	15.7327	15.7170	0.6193	0.8686
18	17	16	2	3	15	0.7408	0.4033	0.2153	17.0402	17.7810	17.4435	17.2556	0.9806	0.8695
19	18	17	2	5	19	0.6539	0.3644	0.0826	10.6339	11.2878	10.9983	10.7165	0.9000	0.8704
20	19	17	2	6	16	0.6641	0.3682	0.3552	7.1234	7.7875	7.4916	7.4785	0.9951	0.8712

Tabela 5.40: Análise SNP 162 - areas

5.2.21 SNP 186

Figura 5.42: Gráficos descritivos dos dos picos e áreas do espectro de massa: histogramas, distribuições acumuladas empíricas. Os dados dos pais estão em destaque.

Figura 5.43: Gráficos de dispersão e genotipagem para as diferentes medidas aplicadas.

p	G_1	G_2	EG_1	EG_2	EF	EQ	Qui	MGE	D_{sep}	L_1	L_2	L_3	$FBST$	$evCr$
2	0	0	0	0	0	2.0406	2.0406	228857.4149	0.0000	2.0406	2.0406	228857.4149	0.3292	0.7969
3	0	0	0	0	18	2.0406	2.0406	228857.4149	10.4033	12.4439	12.4439	228867.8183	0.7500	0.8189
4	1	0	2	0	0	0.8810	0.7413	0.9877	11.2356	12.1167	11.9769	12.2233	0.6000	0.8312
5	1	0	0	0	0	0.2708	0.2260	0.9754	0.0000	0.2708	0.2260	0.9754	0.0000	0.8393
6	1	0	0	0	0	0.0769	0.0662	0.7061	0.0000	0.0769	0.0662	0.7061	0.0000	0.8451
7	1	0	0	0	1	0.0581	0.0511	0.8042	0.0000	0.0581	0.0511	0.8042	0.7500	0.8496
8	1	0	0	0	9	0.1095	0.0975	6.2135	0.3511	0.4606	0.4486	6.5646	0.1118	0.8531
9	1	0	0	0	18	0.1883	0.1696	20.5232	3.4678	3.6560	3.6374	23.9910	0.8000	0.8560
10	1	0	0	0	40	0.2757	0.2508	51.1207	12.4840	12.7597	12.7348	63.6047	0.1062	0.8584
11	2	0	5	0	0	0.0781	0.0684	0.2433	0.8526	0.9307	0.9209	1.0959	0.4970	0.8605
12	2	0	2	0	1	0.0611	0.0537	0.2297	3.7568	3.8179	3.8104	3.9865	0.2865	0.8622
13	2	0	0	0	7	0.0508	0.0444	0.0489	0.8597	0.9106	0.9042	0.9086	0.5000	0.8638
14	2	0	0	0	10	0.0568	0.0498	0.7961	0.8171	0.8739	0.8669	1.6132	0.6000	0.8652
15	2	0	0	0	18	0.0775	0.0685	2.5324	2.0807	2.1582	2.1491	4.6130	0.8750	0.8665
16	2	0	0	0	33	0.1083	0.0963	5.7892	4.1012	4.2095	4.1975	9.8903	0.4545	0.8676
17	2	0	2	0	48	0.1456	0.1303	11.1320	6.5050	6.6506	6.6353	17.6370	0.5048	0.8686
18	3	0	5	0	6	0.0297	0.0257	0.0916	0.9804	1.0101	1.0061	1.0720	0.9069	0.8695
19	3	0	3	0	10	0.0328	0.0287	0.5816	3.0354	3.0682	3.0640	3.6169	0.8500	0.8704
20	3	0	2	0	13	0.0422	0.0369	0.7698	3.8076	3.8499	3.8445	4.5775	0.9691	0.8712

Tabela 5.41: Análise SNP 186 - picos

p	G_1	G_2	EG_1	EG_2	EF	EQ	Qui	MGE	D_{sep}	L_1	L_2	L_3	$FBST$	$euCr$
2	0	0	0	0	0	1.6913	1.6913	79204.4189	0.0000	1.6913	1.6913	79204.4189	0.3284	0.7969
3	0	0	0	0	10	1.6913	1.6913	79204.4189	2.9592	4.6504	4.6504	79207.3781	0.7500	0.8189
4	1	0	4	0	0	0.9944	0.8693	0.8483	15.2929	16.2873	16.1622	16.1412	0.6000	0.8312
5	1	0	0	0	0	0.3957	0.3328	0.9877	0.0000	0.3957	0.3328	0.9877	0.0000	0.8393
6	1	0	0	0	0	0.1274	0.1093	0.8983	0.0000	0.1274	0.1093	0.8983	0.0001	0.8451
7	1	0	0	0	0	0.0515	0.0452	0.3756	0.0000	0.0515	0.0452	0.3756	0.7500	0.8496
8	1	0	0	0	3	0.0601	0.0535	1.4965	0.0390	0.0991	0.0925	1.5355	0.1123	0.8531
9	1	0	0	0	10	0.1055	0.0950	6.4490	0.9864	1.0919	1.0814	7.4354	0.8000	0.8560
10	1	0	1	0	22	0.1663	0.1513	17.0384	3.1210	3.2873	3.2723	20.1594	0.1058	0.8584
11	2	0	5	0	0	0.0840	0.0739	0.1062	4.9983	5.0823	5.0722	5.1045	0.2513	0.8605
12	2	0	4	0	1	0.0660	0.0584	0.1274	5.1439	5.2098	5.2023	5.2713	0.2204	0.8622
13	2	0	1	0	2	0.0504	0.0447	0.1308	3.4651	3.5154	3.5098	3.5959	0.5000	0.8638
14	2	0	0	0	3	0.0447	0.0395	0.0190	2.7443	2.7890	2.7838	2.7633	0.6000	0.8652
15	2	0	0	0	10	0.0472	0.0418	0.4721	0.6011	0.6483	0.6429	1.0732	0.8750	0.8665
16	2	0	0	0	19	0.0598	0.0533	1.6434	1.3167	1.3765	1.3699	2.9600	0.4323	0.8676
17	2	0	1	0	24	0.0799	0.0716	3.4457	3.4200	3.4998	3.4915	6.8657	0.5033	0.8686
18	2	0	5	0	40	0.1055	0.0950	6.4490	6.5717	6.6772	6.6667	13.0206	0.6200	0.8695
19	3	0	4	0	3	0.0253	0.0222	0.0998	3.2989	3.3242	3.3212	3.3988	0.8500	0.8704
20	3	0	4	0	7	0.0267	0.0235	0.2474	3.1210	3.1477	3.1445	3.3684	0.9667	0.8712

Tabela 5.42: Análise SNP 186 - areas

5.2.22 SNP 201

Figura 5.44: Gráficos descritivos dos dos picos e áreas do espectro de massa: histogramas, distribuições acumuladas empíricas. Os dados dos pais estão em destaque.

Figura 5.45: Gráficos de dispersão e genotipagem para as diferentes medidas aplicadas.

p	G_1	G_2	EG_1	EG_2	EF	EQ	Qui	MGE	D_{sep}	L_1	L_2	L_3	$FBST$	$evCr$
2	2	2	2	0	4	4.3872	2.1936	1.0000	0.0000	4.3872	2.1936	1.0000	0.3301	0.7969
3	3	3	2	0	4	2.8933	1.5907	0.9958	3.3406	6.2339	4.9313	4.3365	0.5000	0.8189
4	4	4	4	1	58	4.3872	2.1936	1.0000	2.2542	6.6414	4.4478	3.2542	0.8000	0.8312
5	4	4	5	6	4	1.1425	0.6638	0.0723	24.9680	26.1106	25.6318	25.0403	0.0018	0.8393
6	5	5	2	2	8	1.3667	0.7893	1.4972	19.9488	21.3154	20.7381	21.4460	0.1617	0.8451
7	6	6	2	0	8	0.9317	0.5655	0.9212	8.4959	9.4275	9.0614	9.4170	0.7500	0.8496
8	7	7	2	0	12	1.3065	0.7404	0.2346	3.0102	4.3167	3.7506	3.2448	0.3311	0.8531
9	8	8	2	0	12	1.0082	0.5975	0.3398	1.9424	2.9506	2.5399	2.2822	0.7000	0.8560
10	9	9	2	0	16	1.3622	0.7555	0.5155	1.6015	2.9637	2.3570	2.1170	0.2795	0.8584
11	10	10	2	0	16	1.1255	0.6486	0.7285	1.2703	2.3958	1.9188	1.9987	0.5696	0.8605
12	11	11	4	1	23	1.4723	0.8019	0.8675	0.9257	2.3980	1.7276	1.7932	0.3887	0.8622
13	12	12	5	3	20	1.2726	0.7161	0.8453	2.0806	3.3531	2.7966	2.9259	0.6816	0.8638
14	12	12	7	6	20	1.1099	0.6426	0.7705	5.3680	6.4780	6.0107	6.1385	0.5990	0.8652
15	13	13	5	6	51	1.0949	0.6410	0.0126	9.0047	10.0996	9.6457	9.0172	0.8125	0.8665
16	14	14	3	5	24	1.1903	0.6775	0.6942	8.4400	9.6303	9.1175	9.1342	0.9874	0.8676
17	15	15	2	4	55	1.1558	0.6672	0.1161	7.9792	9.1350	8.6464	8.0954	0.9873	0.8686
18	16	16	2	2	28	1.2548	0.7044	0.6314	7.5918	8.8466	8.2962	8.2232	0.9928	0.8695
19	17	17	2	3	59	1.2082	0.6887	0.1898	6.7191	7.9273	7.4078	6.9089	0.8094	0.8704
20	18	18	4	2	33	1.3179	0.7312	0.6873	4.9570	6.2749	5.6883	5.6443	0.9993	0.8712

Tabela 5.43: Análise SNP 201 - picos

p	G_1	G_2	EG_1	EG_2	EF	EQ	Qui	MGE	D_{sep}	L_1	L_2	L_3	$FBST$	$evCr$
2	2	2	2	0	4	4.8394	2.4197	1.0000	0.0390	4.8785	2.4587	1.0390	0.3302	0.7969
3	3	3	2	0	4	3.0235	1.6618	0.9937	2.7771	5.8006	4.4389	3.7708	0.5000	0.8189
4	4	4	5	2	72	4.8394	2.4197	1.0000	5.6373	10.4768	8.0570	6.6373	0.8000	0.8312
5	4	4	3	6	4	1.0813	0.6367	1.6255	28.8537	29.9349	29.4903	30.4791	0.1358	0.8393
6	5	5	2	2	8	1.3127	0.7616	0.9505	12.1101	13.4228	12.8717	13.0606	0.1986	0.8451
7	6	6	2	0	8	0.9129	0.5561	0.2105	4.1904	5.1033	4.7465	4.4009	0.7500	0.8496
8	7	7	2	0	12	1.2886	0.7314	0.3056	2.0087	3.2973	2.7401	2.3142	0.3326	0.8531
9	8	8	2	0	12	0.9994	0.5929	0.5194	1.1329	2.1323	1.7258	1.6523	0.7000	0.8560
10	9	9	2	1	16	1.3710	0.7599	0.7580	0.9636	2.3346	1.7235	1.7216	0.2831	0.8584
11	10	10	3	1	16	1.1535	0.6635	0.8202	1.1892	2.3427	1.8527	2.0094	0.5411	0.8605
12	11	11	5	2	30	1.5206	0.8279	0.8951	2.3347	3.8553	3.1626	3.2298	0.4692	0.8622
13	11	12	6	5	49	1.2457	0.7170	0.1921	4.4290	5.6747	5.1460	4.6210	0.6377	0.8638
14	12	12	5	6	20	1.1202	0.6480	0.7458	7.2817	8.4019	7.9298	8.0275	0.6071	0.8652
15	13	13	3	6	51	1.0960	0.6417	0.1395	10.2063	11.3022	10.8480	10.3458	0.8125	0.8665
16	14	14	2	6	24	1.1891	0.6771	0.6589	7.7313	8.9204	8.4084	8.3902	0.9914	0.8676
17	15	15	2	4	55	1.1583	0.6686	0.1067	6.2889	7.4472	6.9575	6.3956	0.9904	0.8686
18	16	16	2	3	28	1.2613	0.7080	0.6779	5.2072	6.4685	5.9152	5.8851	0.9972	0.8695
19	17	17	3	2	59	1.2165	0.6932	0.2155	4.3499	5.5664	5.0431	4.5654	0.8000	0.8704
20	18	18	5	2	36	1.3290	0.7373	0.7403	3.5918	4.9208	4.3291	4.3321	0.9996	0.8712

Tabela 5.44: Análise SNP 201 - areas

5.2.23 SNP 204

Figura 5.46: Gráficos descritivos dos dos picos e áreas do espectro de massa: histogramas, distribuições acumuladas empíricas. Os dados dos pais estão em destaque.

Figura 5.47: Gráficos de dispersão e genotipagem para as diferentes medidas aplicadas.

p	G_1	G_2	EG_1	EG_2	EF	EQ	Qui	MGE	D_{sep}	L_1	L_2	L_3	$FBST$	$evCr$
2	0	0	2	0	1	3.0676	3.0676	140560376.9350	0.0000	3.0676	3.0676	140560376.9350	0.3290	0.7969
3	0	0	2	0	16	3.0676	3.0676	140560376.9350	38.4545	41.5221	41.5221	140560415.3895	0.7500	0.8189
4	0	0	7	0	121	3.0676	3.0676	140560376.9350	3.7809	6.8484	6.8484	140560380.7158	0.8000	0.8312
5	1	0	2	0	2	0.4899	0.4119	0.9876	0.0000	0.4899	0.4119	0.9876	0.2252	0.8393
6	1	0	2	0	2	0.2140	0.1865	0.5578	0.0000	0.2140	0.1865	0.5578	0.1419	0.8451
7	1	0	2	0	3	0.1896	0.1686	5.0658	0.0000	0.1896	0.1686	5.0658	0.7500	0.8496
8	1	0	2	0	8	0.2659	0.2388	43.8221	0.4075	0.6734	0.6463	44.2296	0.1777	0.8531
9	1	0	2	0	18	0.3812	0.3453	216.1009	12.8453	13.2265	13.1905	228.9462	0.8000	0.8560
10	1	0	2	0	56	0.5087	0.4645	777.0220	33.7346	34.2433	34.1990	810.7566	0.1745	0.8584
11	1	0	4	0	100	0.6364	0.5852	2231.1105	11.8515	12.4879	12.4367	2242.9620	0.5000	0.8605
12	1	0	7	0	124	0.7582	0.7014	4991.4453	1.3696	2.1278	2.0710	4992.8149	0.1735	0.8622
13	2	0	6	0	6	0.1813	0.1597	0.6622	0.8067	0.9879	0.9664	1.4688	0.5000	0.8638
14	2	0	2	0	12	0.1876	0.1665	4.2840	1.0510	1.2386	1.2175	5.3350	0.6019	0.8652
15	2	0	2	0	20	0.2175	0.1939	12.5620	9.2364	9.4539	9.4303	21.7984	0.8750	0.8665
16	2	0	2	0	40	0.2623	0.2350	34.1704	20.9168	21.1791	21.1518	55.0872	0.7818	0.8676
17	2	0	2	0	69	0.3166	0.2849	76.4552	21.0066	21.3232	21.2915	97.4618	0.5306	0.8686
18	2	0	3	0	96	0.3761	0.3399	147.6942	16.0852	16.4612	16.4251	163.7793	0.8963	0.8695
19	2	0	5	0	122	0.4385	0.3981	295.3364	9.4177	9.8561	9.8157	304.7541	0.9000	0.8704
20	2	0	7	0	134	0.5022	0.4578	551.9257	10.5131	11.0153	10.9709	562.4389	0.9506	0.8712

Tabela 5.45: Análise SNP 204 - picos

p	G_1	G_2	EG_1	EG_2	EF	EQ	Qui	MGE	D_{sep}	L_1	L_2	L_3	$FBST$	$evCr$
2	0	0	2	0	1	4.9600	4.9600	78843438063.9776	0.1543	5.1144	5.1144	78843438064.1319	0.3299	0.7969
3	0	0	2	0	113	4.9600	4.9600	78843438063.9776	15.6493	20.6093	20.6093	78843438079.6269	0.7500	0.8189
4	1	0	6	0	1	0.8001	0.6804	0.9612	0.0772	0.8772	0.7576	1.0383	0.6000	0.8312
5	1	0	2	0	2	0.3459	0.3171	3.8127	0.0000	0.3459	0.3171	3.8127	0.2491	0.8393
6	1	0	2	0	2	0.3950	0.3611	132.8863	2.3377	2.7327	2.6988	135.2240	0.1653	0.8451
7	1	0	2	0	14	0.5923	0.5322	684.3066	11.3541	11.9463	11.8863	695.6607	0.7500	0.8496
8	1	0	2	0	61	0.8216	0.7398	3016.5175	25.5594	26.3811	26.2992	3042.0769	0.2123	0.8531
9	1	0	2	0	115	1.0493	0.9493	6372.1555	10.0637	11.1129	11.0130	6382.2192	0.8000	0.8560
10	1	0	3	0	128	1.2643	1.1526	18789.3958	5.2778	6.5421	6.4303	18794.6736	0.1832	0.8584
11	1	0	7	0	136	1.4629	1.3420	29695.8375	2.6850	4.1480	4.0270	29698.5225	0.7500	0.8605
12	2	0	6	0	23	0.2592	0.2253	5.7644	9.5479	9.8071	9.7732	15.3123	0.6052	0.8622
13	2	0	2	0	57	0.3462	0.3019	26.0007	12.5292	12.8754	12.8311	38.5298	0.6032	0.8638
14	2	0	2	0	84	0.4553	0.3996	108.0175	12.6183	13.0736	13.0179	120.6358	0.6585	0.8652
15	2	0	2	0	119	0.5772	0.5102	392.8273	3.5917	4.1688	4.1019	396.4190	0.8750	0.8665
16	2	0	2	0	137	0.7047	0.6272	1220.6843	1.7931	2.4978	2.4203	1222.4773	0.8725	0.8676
17	2	0	5	0	153	0.8333	0.7462	2820.2968	7.3610	8.1944	8.1072	2827.6578	0.6515	0.8686
18	2	0	7	0	177	0.9591	0.8636	6292.9427	9.6132	10.5723	10.4768	6302.5559	0.9352	0.8695
19	3	0	6	0	76	0.2939	0.2546	34.2826	11.0492	11.3431	11.3038	45.3318	0.9000	0.8704
20	3	0	6	0	103	0.3675	0.3203	87.2792	8.2523	8.6198	8.5726	95.5315	0.9970	0.8712

Tabela 5.46: Análise SNP 204 - areas

5.2.24 SNP 225

Figura 5.48: Gráficos descritivos dos dos picos e áreas do espectro de massa: histogramas, distribuições acumuladas empíricas. Os dados dos pais estão em destaque.

Figura 5.49: Gráficos de dispersão e genotipagem para as diferentes medidas aplicadas.

p	G_1	G_2	EG_1	EG_2	EF	EQ	Qui	MGE	D_{sep}	L_1	L_2	L_3	$FBST$	$evCr$
2	1	0	0	1	0	7.8118	7.0582	683351172.2113	0.6242	8.4360	7.6824	683351172.8355	0.3334	0.7969
3	1	0	0	1	0	6.1218	5.2916	0.3934	0.4161	6.5380	5.7078	0.8095	0.5000	0.8189
4	1	0	0	1	0	3.8628	3.5824	22.9686	4.9936	8.8564	8.5760	27.9622	0.6000	0.8312
5	1	0	4	1	10	2.2396	2.0738	4.3100	9.2382	11.4778	11.3119	13.5481	0.3438	0.8393
6	2	0	0	1	0	2.1662	1.8598	0.4017	3.5371	5.7034	5.3969	3.9388	0.3332	0.8451
7	2	0	0	1	0	2.2571	1.9287	0.5296	16.4967	18.7539	18.4255	17.0264	0.7499	0.8496
8	2	0	0	1	8	2.1226	1.8164	0.9055	0.9412	3.0637	2.7575	1.8467	0.3422	0.8531
9	3	0	1	1	0	2.3029	1.9466	0.7676	0.1597	2.4626	2.1063	0.9273	0.7999	0.8560
10	3	0	0	1	1	2.1186	1.8355	1.4550	0.3745	2.4931	2.2101	1.8296	0.4652	0.8584
11	3	0	0	1	8	2.0927	1.8384	6.5827	6.6955	8.7882	8.5339	13.2782	0.5002	0.8605
12	4	0	5	1	0	2.6008	2.2254	0.9139	11.5896	14.1904	13.8150	12.5035	0.9833	0.8622
13	4	0	0	1	2	2.4498	2.1211	0.0811	5.3165	7.7663	7.4376	5.3976	0.8514	0.8638
14	4	0	0	1	8	2.2583	1.9753	0.7049	5.8717	8.1300	7.8469	6.5765	0.7875	0.8652
15	4	0	4	1	10	2.0051	1.7577	1.5000	5.3843	7.3894	7.1420	6.8843	0.7514	0.8665
16	5	0	2	1	2	2.1954	1.8850	0.2732	7.7129	9.9083	9.5979	7.9861	0.9999	0.8676
17	5	0	0	1	5	2.1923	1.8939	0.4051	7.4571	9.6494	9.3510	7.8621	0.9992	0.8686
18	5	0	3	1	9	2.0952	1.8189	0.3667	3.1788	5.2740	4.9977	3.5455	1.0000	0.8695
19	6	0	5	1	2	2.2156	1.9003	0.0327	2.6259	4.8415	4.5262	2.6586	0.9000	0.8704
20	6	0	1	1	4	2.1806	1.8883	0.5784	4.9624	7.1430	6.8507	5.5408	1.0000	0.8712

Tabela 5.47: Análise SNP 225 - picos

p	G_1	G_2	EG_1	EG_2	EF	EQ	Qui	MGE	D_{sep}	L_1	L_2	L_3	$FBST$	$evCr$
2	1	0	0	1	0	6.9358	6.2272	143039646.9851	0.0000	6.9358	6.2272	143039646.9851	0.3334	0.7969
3	1	0	0	1	0	6.1230	5.2577	0.0022	5.0976	11.2207	10.3554	5.0998	0.5000	0.8189
4	1	0	0	1	1	3.9237	3.6013	9.6327	0.9558	4.8795	4.5571	10.5885	0.6000	0.8312
5	2	0	1	1	0	3.3658	3.0592	10.1241	20.2865	23.6523	23.3457	30.4106	0.0243	0.8393
6	2	0	0	1	0	2.7086	2.4183	2.1860	3.6123	6.3208	6.0306	5.7982	0.2795	0.8451
7	2	0	1	1	1	2.3907	2.1082	2.1370	10.5332	12.9239	12.6414	12.6701	0.6250	0.8496
8	3	0	1	1	0	2.5919	2.2374	0.6717	8.9948	11.5868	11.2322	9.6665	0.3737	0.8531
9	3	0	0	1	0	2.4075	2.0584	0.8478	2.6569	5.0644	4.7153	3.5047	0.7990	0.8560
10	3	0	2	1	1	2.1574	1.8637	0.5754	0.2809	2.4382	2.1446	0.8563	0.4560	0.8584
11	4	0	1	1	0	2.1406	1.8578	0.5591	2.5798	4.7204	4.4376	3.1389	0.5000	0.8605
12	4	0	0	1	1	2.2637	1.9654	0.7863	6.1640	8.4277	8.1294	6.9503	0.9291	0.8622
13	4	0	4	1	1	2.3663	2.0601	0.1575	6.6814	9.0477	8.7414	6.8389	0.8784	0.8638
14	5	0	1	1	0	2.3738	2.0647	0.0542	3.9790	6.3528	6.0437	4.0332	0.6361	0.8652
15	5	0	2	1	1	2.3167	2.0221	0.3642	9.0070	11.3237	11.0291	9.3712	0.8125	0.8665
16	5	0	5	1	1	2.3202	2.0241	0.1625	4.1655	6.4857	6.1896	4.3280	0.9997	0.8676
17	6	0	2	1	1	2.2805	1.9773	0.2074	5.4307	7.7112	7.4081	5.6381	0.9998	0.8686
18	6	0	2	1	1	2.2853	1.9914	0.0847	5.0533	7.3387	7.0447	5.1380	1.0000	0.8695
19	6	0	6	1	3	2.2937	2.0053	0.0132	3.2531	5.5469	5.2585	3.2664	0.8000	0.8704
20	7	0	2	1	1	2.3246	2.0263	0.2145	3.5540	5.8787	5.5804	3.7686	1.0000	0.8712

Tabela 5.48: Análise SNP 225 - areas

5.2.25 SNP 235

Figura 5.50: Gráficos descritivos dos dos picos e áreas do espectro de massa: histogramas, distribuições acumuladas empíricas. Os dados dos pais estão em destaque.

Figura 5.51: Gráficos de dispersão e genotipagem para as diferentes medidas aplicadas.

p	G_1	G_2	EG_1	EG_2	EF	EQ	Qui	MGE	D_{sep}	L_1	L_2	L_3	$FBST$	$euCr$
2	0	0	0	0	0	1.6997	1.6997	94234.9521	0.0000	1.6997	1.6997	94234.9521	0.3295	0.7969
3	0	0	0	0	2	1.6997	1.6997	94234.9521	0.1829	1.8826	1.8826	94235.1350	0.7500	0.8189
4	0	0	0	0	75	1.6997	1.6997	94234.9521	37.3457	39.0454	39.0454	94272.2977	0.8000	0.8312
5	0	1	0	1	0	0.3359	0.2801	0.9878	0.0000	0.3359	0.2801	0.9878	0.0000	0.8393
6	0	1	0	0	0	0.0799	0.0686	0.8588	0.0000	0.0799	0.0686	0.8588	0.0000	0.8451
7	0	1	0	0	0	0.0154	0.0136	0.1510	0.0000	0.0154	0.0136	0.1510	0.7500	0.8496
8	0	1	0	0	0	0.0317	0.0283	2.3415	0.0386	0.0703	0.0669	2.3801	0.1123	0.8531
9	0	1	0	0	2	0.0828	0.0746	8.8472	0.0610	0.1437	0.1356	8.9082	0.8000	0.8560
10	0	1	0	0	14	0.1478	0.1345	22.6077	12.3457	12.4935	12.4801	34.9533	0.1031	0.8584
11	0	1	0	0	50	0.2170	0.1990	47.5973	28.7318	28.9488	28.9308	76.3290	0.5000	0.8605
12	0	1	0	4	75	0.2856	0.2637	88.1073	12.4486	12.7342	12.7123	100.5559	0.0969	0.8622
13	0	2	0	4	0	0.0322	0.0281	0.6043	0.0809	0.1131	0.1090	0.6853	0.5000	0.8638
14	0	2	0	1	1	0.0154	0.0136	0.1510	0.0450	0.0603	0.0586	0.1960	0.6000	0.8652
15	0	2	0	1	2	0.0172	0.0153	0.7583	0.0366	0.0538	0.0518	0.7949	0.8750	0.8665
16	0	2	0	1	5	0.0317	0.0283	2.3415	3.7809	3.8126	3.8092	6.1224	0.4044	0.8676
17	0	2	0	0	20	0.0546	0.0489	4.9117	10.0514	10.1060	10.1004	14.9631	0.5036	0.8686
18	0	2	0	0	42	0.0828	0.0746	8.8472	17.5583	17.6411	17.6329	26.4055	0.5978	0.8695
19	0	2	0	1	65	0.1143	0.1035	14.5848	5.4511	5.5654	5.5546	20.0359	0.9000	0.8704
20	0	2	0	4	77	0.1478	0.1345	22.6077	7.4691	7.6169	7.6036	30.0768	0.8073	0.8712

Tabela 5.49: Análise SNP 235 - picos

p	G_1	G_2	EG_1	EG_2	EF	EQ	Qui	MGE	D_{sep}	L_1	L_2	L_3	$FBST$	$euCr$
2	0	0	0	0	0	1.1624	1.1624	14240.4426	0.0000	1.1624	1.1624	14240.4426	0.3288	0.7969
3	0	0	0	0	0	1.1624	1.1624	14240.4426	0.0000	1.1624	1.1624	14240.4426	0.7500	0.8189
4	0	0	0	0	1	1.1624	1.1624	14240.4426	52.1605	53.3229	53.3229	14292.6031	0.8000	0.8312
5	0	1	0	3	0	0.6313	0.5401	0.9905	8.8889	9.5202	9.4290	9.8794	0.0000	0.8393
6	0	1	0	0	0	0.2476	0.2123	0.9787	0.0000	0.2476	0.2123	0.9787	0.0001	0.8451
7	0	1	0	0	0	0.0824	0.0722	0.8717	0.0000	0.0824	0.0722	0.8717	0.7500	0.8496
8	0	1	0	0	0	0.0232	0.0207	0.4950	0.0000	0.0232	0.0207	0.4950	0.1122	0.8531
9	0	1	0	0	0	0.0155	0.0140	0.4882	0.0000	0.0155	0.0140	0.4882	0.8000	0.8560
10	0	1	0	0	1	0.0335	0.0305	2.5677	1.1111	1.1446	1.1416	3.6788	0.1058	0.8584
11	0	1	0	0	4	0.0642	0.0589	6.3443	0.6762	0.7404	0.7351	7.0205	0.5000	0.8605
12	0	1	0	1	14	0.1008	0.0931	12.4664	17.3868	17.4877	17.4800	29.8532	0.0973	0.8622
13	0	1	0	3	39	0.1396	0.1297	21.5677	23.3854	23.5250	23.5151	44.9531	0.1384	0.8638
14	0	2	0	5	0	0.0613	0.0550	0.3559	11.2317	11.2929	11.2867	11.5876	0.5958	0.8652
15	0	2	0	3	0	0.0394	0.0353	0.5414	2.9630	3.0024	2.9983	3.5044	0.8745	0.8665
16	0	2	0	2	1	0.0226	0.0203	0.4338	0.3954	0.4181	0.4157	0.8293	0.4320	0.8676
17	0	2	0	1	2	0.0149	0.0134	0.1066	0.1108	0.1257	0.1242	0.2174	0.5031	0.8686
18	0	2	0	0	4	0.0155	0.0140	0.4882	0.1431	0.1586	0.1571	0.6313	0.5970	0.8695
19	0	2	0	0	7	0.0223	0.0202	1.3553	1.6588	1.6811	1.6791	3.0141	0.9000	0.8704
20	0	2	0	1	14	0.0335	0.0305	2.5677	10.4321	10.4656	10.4626	12.9998	0.7956	0.8712

Tabela 5.50: Análise SNP 235 - areas

5.2.26 SNP 237

Figura 5.52: Gráficos descritivos dos dos picos e áreas do espectro de massa: histogramas, distribuições acumuladas empíricas. Os dados dos pais estão em destaque.

Figura 5.53: Gráficos de dispersão e genotipagem para as diferentes medidas aplicadas.

p	G_1	G_2	EG_1	EG_2	EF	EQ	Qui	MGE	D_{sep}	L_1	L_2	L_3	$FBST$	$evCr$
2	0	0	0	0	0	1.9262	1.9262	597912.8574	0.0000	1.9262	1.9262	597912.8574	0.3290	0.7969
3	0	0	0	0	7	1.9262	1.9262	597912.8574	6.5844	8.5105	8.5105	597919.4417	0.7500	0.8189
4	0	1	0	0	0	1.0536	0.8926	0.9742	44.4444	45.4980	45.3371	45.4187	0.6000	0.8312
5	0	1	0	0	0	0.3481	0.2938	0.9690	0.0000	0.3481	0.2938	0.9690	0.0000	0.8393
6	0	1	0	0	0	0.1000	0.0889	0.5855	0.0000	0.1000	0.0889	0.5855	0.0001	0.8451
7	0	1	0	0	0	0.0479	0.0448	1.7630	0.0900	0.1379	0.1348	1.8530	0.7500	0.8496
8	0	1	0	0	1	0.0761	0.0697	9.4581	0.1543	0.2305	0.2240	9.6125	0.1133	0.8531
9	0	1	0	0	7	0.1382	0.1262	32.2041	2.1948	2.3330	2.3210	34.3989	0.8000	0.8560
10	0	1	0	0	29	0.2136	0.1959	83.5510	27.7778	27.9914	27.9736	111.3288	0.1100	0.8584
11	0	1	0	3	58	0.2925	0.2696	181.9374	12.6972	12.9897	12.9668	194.6346	0.5000	0.8605
12	0	2	0	0	0	0.0790	0.0717	0.2524	14.8148	14.8939	14.8865	15.0672	0.3138	0.8622
13	0	2	0	0	0	0.0505	0.0463	0.3084	0.1309	0.1814	0.1773	0.4393	0.5000	0.8638
14	0	2	0	0	4	0.0424	0.0395	1.9341	0.6335	0.6759	0.6730	2.5676	0.6000	0.8652
15	0	2	0	0	7	0.0518	0.0478	4.5938	2.0576	2.1094	2.1054	6.6514	0.8750	0.8665
16	0	2	0	0	17	0.0729	0.0668	10.0733	7.9958	8.0686	8.0625	18.0691	0.4674	0.8676
17	0	2	0	1	36	0.1018	0.0928	17.1191	11.8230	11.9248	11.9158	28.9421	0.5077	0.8686
18	0	2	0	2	53	0.1357	0.1237	27.1543	8.0899	8.2256	8.2136	35.2441	0.7266	0.8695
19	0	3	0	5	2	0.0397	0.0367	0.5323	7.6743	7.7141	7.7110	8.2066	0.8500	0.8704
20	0	3	0	0	4	0.0360	0.0335	1.1469	15.1235	15.1595	15.1570	16.2703	0.9850	0.8712

Tabela 5.51: Análise SNP 237 - picos

p	G_1	G_2	EG_1	EG_2	EF	EQ	Qui	MGE	D_{sep}	L_1	L_2	L_3	$FBST$	$euCr$
2	0	0	0	0	0	1.2051	1.2051	32145.2722	0.0000	1.2051	1.2051	32145.2722	0.3290	0.7969
3	0	0	0	0	0	1.2051	1.2051	32145.2722	0.0000	1.2051	1.2051	32145.2722	0.7500	0.8189
4	0	0	0	1	23	1.2051	1.2051	32145.2722	15.1235	16.3286	16.3286	32160.3956	0.8000	0.8312
5	0	1	0	0	0	0.6642	0.5815	0.9467	12.0988	12.7630	12.6803	13.0454	0.0000	0.8393
6	0	1	0	0	0	0.2923	0.2524	0.9740	0.0229	0.3152	0.2752	0.9969	0.0001	0.8451
7	0	1	0	0	0	0.1121	0.0990	0.8515	0.0000	0.1121	0.0990	0.8515	0.7500	0.8496
8	0	1	0	0	0	0.0433	0.0393	0.3674	0.0000	0.0433	0.0393	0.3674	0.1132	0.8531
9	0	1	0	0	0	0.0305	0.0281	0.9835	0.1084	0.1388	0.1365	1.0919	0.8000	0.8560
10	0	1	0	0	0	0.0451	0.0414	3.5458	0.0000	0.0451	0.0414	3.5458	0.1132	0.8584
11	0	1	0	0	0	0.0737	0.0679	8.8355	1.7957	1.8694	1.8637	10.6312	0.5000	0.8605
12	0	1	0	1	23	0.1093	0.1012	17.8042	5.0412	5.1504	5.1424	22.8453	0.1175	0.8622
13	0	1	0	3	39	0.1477	0.1375	31.6554	13.5005	13.6482	13.6379	45.1559	0.1578	0.8638
14	0	2	0	3	0	0.0610	0.0555	0.3271	8.3216	8.3827	8.3771	8.6488	0.4218	0.8652
15	0	2	0	0	0	0.0448	0.0408	0.0484	4.0329	4.0777	4.0737	4.0813	0.8125	0.8665
16	0	2	0	0	2	0.0328	0.0301	0.1821	2.5704	2.6032	2.6005	2.7525	0.5018	0.8676
17	0	2	0	0	2	0.0272	0.0249	0.3558	0.8607	0.8878	0.8856	1.2165	0.5059	0.8686
18	0	2	0	0	4	0.0268	0.0246	0.9780	1.1050	1.1318	1.1296	2.0831	0.6819	0.8695
19	0	2	0	0	10	0.0325	0.0298	2.2151	5.0603	5.0928	5.0901	7.2754	0.9000	0.8704
20	0	2	0	1	23	0.0429	0.0394	3.7623	3.0247	3.0676	3.0641	6.7870	0.8353	0.8712

Tabela 5.52: Análise SNP 237 - areas

5.2.27 SNP 241

Figura 5.54: Gráficos descritivos dos dos picos e áreas do espectro de massa: histogramas, distribuições acumuladas empíricas. Os dados dos pais estão em destaque.

Figura 5.55: Gráficos de dispersão e genotipagem para as diferentes medidas aplicadas.

p	G_1	G_2	EG_1	EG_2	EF	EQ	Qui	MGE	D_{sep}	L_1	L_2	L_3	$FBST$	$evCr$
2	0	0	0	0	0	1.5508	1.5508	150758.9743	0.0000	1.5508	1.5508	150758.9743	0.3291	0.7969
3	0	0	0	0	1	1.5508	1.5508	150758.9743	0.0000	1.5508	1.5508	150758.9743	0.7500	0.8189
4	0	0	0	2	41	1.5508	1.5508	150758.9743	60.4938	62.0447	62.0447	150819.4681	0.8000	0.8312
5	0	1	0	1	0	0.6452	0.5565	0.9906	8.8889	9.5340	9.4454	9.8795	0.0000	0.8393
6	0	1	0	0	0	0.2483	0.2149	0.9776	0.0229	0.2712	0.2378	1.0005	0.0039	0.8451
7	0	1	0	0	0	0.0856	0.0759	0.8396	0.0900	0.1756	0.1659	0.9296	0.7500	0.8496
8	0	1	0	0	0	0.0380	0.0342	0.2782	0.0000	0.0380	0.0342	0.2782	0.1218	0.8531
9	0	1	0	0	1	0.0456	0.0414	1.6273	0.0000	0.0456	0.0414	1.6273	0.8000	0.8560
10	0	1	0	0	5	0.0806	0.0736	6.4722	1.2346	1.3151	1.3082	7.7067	0.1303	0.8584
11	0	1	0	1	18	0.1277	0.1172	15.2363	5.5004	5.6281	5.6175	20.7367	0.5000	0.8605
12	0	1	0	2	41	0.1800	0.1663	32.7493	20.2103	20.3903	20.3766	52.9596	0.1522	0.8622
13	0	1	0	4	66	0.2335	0.2168	57.4399	11.5365	11.7700	11.7533	68.9764	0.1732	0.8638
14	0	2	0	5	0	0.0510	0.0457	0.4944	4.1482	4.1992	4.1939	4.6426	0.5999	0.8652
15	0	2	0	1	1	0.0348	0.0312	0.3524	2.9630	2.9977	2.9942	3.3154	0.8125	0.8665
16	0	2	0	0	1	0.0274	0.0246	0.0460	0.8729	0.9002	0.8975	0.9189	0.5994	0.8676
17	0	2	0	0	6	0.0297	0.0266	0.4188	1.9223	1.9521	1.9490	2.3411	0.5142	0.8686
18	0	2	0	1	16	0.0394	0.0355	1.6127	3.3684	3.4078	3.4039	4.9811	0.7821	0.8695
19	0	2	0	1	27	0.0556	0.0503	3.5256	8.0539	8.1095	8.1043	11.5796	0.9000	0.8704
20	0	2	0	2	41	0.0758	0.0690	6.1002	12.1605	12.2363	12.2294	18.2607	0.8998	0.8712

Tabela 5.53: Análise SNP 241 - picos

p	G_1	G_2	EG_1	EG_2	EF	EQ	Qui	MGE	D_{sep}	L_1	L_2	L_3	$FBST$	$euCr$
2	0	0	0	0	0	1.1765	1.1765	33462.9534	0.0000	1.1765	1.1765	33462.9534	0.3303	0.7969
3	0	0	0	0	0	1.1765	1.1765	33462.9534	0.0000	1.1765	1.1765	33462.9534	0.7500	0.8189
4	0	0	0	1	12	1.1765	1.1765	33462.9534	11.1111	12.2876	12.2876	33474.0645	0.8000	0.8312
5	0	0	0	2	65	1.1765	1.1765	33462.9534	41.7284	42.9049	42.9049	33504.6818	0.5000	0.8393
6	0	1	0	1	0	0.3983	0.3525	0.9729	6.6072	7.0055	6.9597	7.5802	0.0000	0.8451
7	0	1	0	0	0	0.1739	0.1538	0.9535	0.0900	0.2639	0.2438	1.0435	0.7500	0.8496
8	0	1	0	0	0	0.0728	0.0656	0.7972	0.0386	0.1114	0.1042	0.8358	0.1201	0.8531
9	0	1	0	0	0	0.0368	0.0336	0.3520	0.1084	0.1452	0.1420	0.4604	0.8000	0.8560
10	0	1	0	0	1	0.0357	0.0327	0.6586	0.0000	0.0357	0.0327	0.6586	0.1324	0.8584
11	0	1	0	0	4	0.0535	0.0492	2.9316	0.7272	0.7807	0.7764	3.6588	0.5000	0.8605
12	0	1	0	1	12	0.0814	0.0753	7.1156	3.7037	3.7851	3.7790	10.8193	0.1509	0.8622
13	0	1	0	1	28	0.1144	0.1063	14.0442	11.5000	11.6143	11.6063	25.5442	0.1721	0.8638
14	0	1	0	1	42	0.1494	0.1396	24.6096	14.0014	14.1508	14.1410	38.6109	0.6000	0.8652
15	0	1	0	2	65	0.1848	0.1733	37.1559	13.9918	14.1766	14.1651	51.1476	0.8750	0.8665
16	0	1	0	5	79	0.2194	0.2066	56.5491	3.0961	3.3155	3.3026	59.6452	0.2136	0.8676
17	0	2	0	4	1	0.0301	0.0275	0.2497	3.6150	3.6451	3.6425	3.8647	0.5191	0.8686
18	0	2	0	1	1	0.0258	0.0236	0.0196	3.0907	3.1164	3.1142	3.1103	0.7738	0.8695
19	0	2	0	0	6	0.0260	0.0236	0.1629	1.4362	1.4622	1.4598	1.5990	0.9000	0.8704
20	0	2	0	1	12	0.0300	0.0274	0.7297	2.2222	2.2523	2.2496	2.9519	0.9028	0.8712

Tabela 5.54: Análise SNP 241 - areas

Capítulo 6

Considerações finais e próximos passos

Neste capítulo finalizamos o trabalho comparando com um método recente e com uma abordagem diferente para a resolução do mesmo problema, descrevemos o fluxo da análise desenvolvida e finalmente listamos as contribuições, limitações e finalmente as oportunidades de futuros estudos que ficaram em aberto.

6.1 Comparação com o método SuperMASSA

O método SuperMassa [9] é baseado num modelo de redes bayesianas e envolve a inferência de um ponto de máximo a posteriori. Maiores detalhes podem ser encontrados no endereço <http://statgen.esalq.usp.br/SuperMASSA/>.

Os dados de cana-de-açúcar utilizados neste trabalho são os mesmos utilizados no artigo do método SuperMASSA. Na Tabela 6.1, observamos os resultados apresentados para a estimativa de ploidia e o número de agrupamentos estimado pelos dois métodos. As linhas Expert correspondem à avaliação de um pesquisador a respeito dos dados, a linha SuperMASSA mostra os resultados obtidos usando apenas os dados a respeito dos picos apontados pelo espectômetro de massa. As linhas Urnas (L_1 - picos) e Urnas (L_1 - áreas) correspondem ao método deste trabalho usando tanto os dados de picos quanto de áreas do experimento.

Observamos que os resultados apresentados não são muito diferentes, porém o método SuperMASSA assume em seu modelo apenas ploidias pares, uma limitação que pode ser imposta ao método das urnas como restrição. O número de agrupamentos estimados foram praticamente idênticos usando os dados de área, o que pode ser uma evidência de que estes dados são mais coerentes para representar fisicamente as massas dos haplótipos contidos no genótipo dos SNP.

6.2 Contribuições

1. Caracterização do problema de genotipagem de poliplóides com o problema de estimação de capacidades de urnas e bolas com restrições físicas.
2. Método de simulação de cruzamentos e de níveis de expressão gênica de poliplóides, de implementação simples e disponível em ambiente de código aberto.
3. Desenvolvimento de um método de estimação de ploidia e genotipagem:
 - (a) flexível e independente de modelos paramétricos;
 - (b) de fácil aplicação, computacionalmente rápido e eficiente para aplicação em grandes bases de dados;
 - (c) empiricamente consistente e dependente da qualidade das medidas experimentais;
 - (d) usa uma transformação que reduz a dimensionalidade e descarta a necessidade de modelos computacionalmente complexos;
4. Métodos gráficos simples e coerentes para descrever os dados e os resultados do método;

		SNPs (Cana-de-açúcar)													
		004	005	013	037	041	045	048	050	060	065	077	079	082	
ploidia	Expert	10	6	12	10	8	8	6	10	10	8	6	6	8	
	SuperMASSA	8	6	14	10	8	8	6	10	14	8	6	6	8	
	Urnas (L_1 - picos)	7	6	13	10	8	8	7	9	10	7	6	6	9	
	Urnas (L_1 - áreas)	6	5	11	2	10	10	6	4	11	8	6	7	7	
	Expert	2	2	2	2	2	2	2	4	3	2	2	2	2	
clusters	SuperMASSA	2	2	2	2	2	2	2	4	4	2	2	2	2	
	Urnas (L_1 - picos)	2	2	4	2	2	2	2	4	3	2	2	2	2	
	Urnas (L_1 - áreas)	2	2	4	1	2	2	2	2	3	2	2	2	2	

		SNPs (Cana-de-açúcar)													
ploidia	Expert	088	114	117	122	136	151	162	186	201	204	225	235	237	241
	SuperMASSA	6	6	10	12	12	6	8	8	10	8	10	8	8	8
	Urnas (L_1 - picos)	6	6	10	12	12	6	8	6	10	6	10	8	8	8
	Urnas (L_1 - áreas)	7	11	14	12	15	5	7	7	11	7	9	7	6	8
	Expert	2	2	3	2	2	2	2	2	3	2	4	2	2	2
clusters	SuperMASSA	2	4	3	2	2	2	2	2	3	2	4	2	2	2
	Urnas (L_1 - picos)	2	2	3	2	2	2	2	2	3	2	4	2	2	2
	Urnas (L_1 - áreas)	2	4	3	2	2	2	2	2	3	2	4	2	2	2

Tabela 6.1: Comparação de métodos SuperMASSA vs. Método de Urnas.

5. Desenvolvimento e aplicação do FBST para o problema do Equilíbrio de Hardy-Weinberg em poliplóides dialélicos;
6. Implementação de simulações em diferentes condições experimentais e aplicação em dados reais com resultados coerentes e consistentes.

6.3 Limitações

Apesar da simplicidade do método de genotipagem proposto, ele possui limitações:

1. Nesta abordagem do problema, a ploidia é uma característica do loco genético em questão e não do indivíduo;
2. Medidas repetidas dos pais com mesma variabilidade e mesma variabilidade dos filhos;
3. Simetria em torno dos pontos de aglomeração;
4. Equivariância entre dados de agrupamentos diferentes;
5. Estimativa da probabilidade de acerto na genotipagem altamente dependente do tamanho amostral;

6.4 Oportunidades

A flexibilidade e limitações do trabalho desenvolvido nesta tese abre oportunidade para estudar.

1. Extensão do método com modelos paramétricos, estimação de erros de classificação e análise de consistência. Estudo do comportamento dos resultados sob distribuições paramétricas ou não-paramétricas unidimensionais com diferentes simetrias;
2. Experimentos que exigem pré-processamento de dados como as plataformas Affymetrix e Illumina. No início do desenvolvimento deste trabalho a idéia inicial era utilizar como referência dados provenientes destas plataformas e replicar o método CRLMM descrito em [Carvalho2006] (e citado no Apêndice D) em dados públicos do projeto HapMap. Nos Apêndices G e H descrevemos alguns modelos necessários no contexto de plataformas Affymetrix e Illumina, onde é necessário um pré-processamento de dados de expressão para a avaliação de análises estatísticas devido aos erros experimentais amplamente conhecidos nesses experimentos.
3. Diferentes arranjos experimentais podem ser utilizados e simulações podem ser avaliadas adicionando ou mudando restrições experimentais;
4. Correlações entre SNPs podem ser consideradas adicionando restrições ao classificador;
5. Extensão do método para dimensões superiores (multialelia);
6. Medidas repetidas com variabilidades diferentes;
7. Agrupamentos com variabilidade diferentes;

Referências Bibliográficas

- [Adham2001] Adham, S. A. and Walker, S. G. (2001). A multivariate Gompertz-type distribution. *Journal of Applied Statistics* 28, 1051–1065.
- [Andrell2004] Andell, Y., Correll, D., Oeth, P. and Jurinke, C. (2004) Multiplexed gene expression analysis using competitive PCR and MassARRAY. Sequenom Application Note. Available at: www.sequenom.com
- [Arellano85] Arellano-Vale, R. B. and Genton, M. G. (1985) On fundamental skew distributions. *Journal of Multivariate Analysis*, v. 96, p. 93-116, 2005. AZZALINI, A. A class of distributions which includes the normal ones. *Scandinavian Journal Statistics*, v. 12, p. 171-178.
- [Arellano2007] Arellano, J. M. (1999). Evidence and Credibility: Full Bayesian Significance Test for Precise Hypotheses. *Entropy Journal*, 1, 69–80.
- [Arnold, 1997] Arnold, M.L. Natural hybridization and evolution. 1997. Oxford Univ. Press. New York.
- [Aultchenko2007] Aultchenko, Y. S., Koning, D. J. & Haley, C. (2007), Genomewide rapid association using mixed model and regression: a fast and simple method for genomewide pedigree-based quantitative trait loci association analysis. *Genetics* 177, p 577-585.
- [Azzalini99] Azzalini, A. and Capitanio, A. (1999) Statistical Applications of the Multivariate Skew Normal Distributions. *Journal Royal Statistics Society*, v. 61, p. 579-602.
- [Azzalini96] Azzalini, A. and Dalla-valle, A. (1996) The multivariate skew-normal distribution. *Biometrika*. v. 83, n. 4, p. 715-726.
- [Balagurunathan2002] Balagurunathan Y. et al. Simulation of cDNA microarrays via a parameterized random signal model. *Journal of Biomedical Optics*. 7:507-523, 2002.
- [Banfield93] Banfield, J. D. and Raftery, A. E. (1993), “Model-based Gaussian and non- Gaussian clustering,” *Biometrics*, 49, 803–821.
- [Bansal2002] Bansal, A., et al. (2002) Association testing by DNA pooling: an effective initial screen. *Proceedings of the National Academy of Sciences of the USA* 99, 16871–16874.
- [Beaulieu2003] Beaulieu, M. and Kosman, D. (2003) Molecular haplotyping combining a novel AS-PCR technique with base-specific cleavage and mass spectrometry. In: *SNPs, Haplotypes and Cancer: Application in Molecular Epidemiology*. Conference Proceedings, Key Biscayne, Florida, September 13–17, A60.
- [Bevan2001] Bevan, M., et al. (2001) Sequence and analysis of the Arabidopsis genome. *Current Opinions in Plant Biology* 4, 105–110.
- [Bolstad2003] Bolstad, B. M. et al. (2003) "A comparison of normalization methods for high density oligonucleotide array data based on variance and bias". *Bioinformatics*, 19 (2):185-193
- [Bolstad2004] Bolstad, B. M. (2004), Low Level Analysis of High-density oligonucleotide array data: Background, normalization and summarization. In PhD Thesis. University of California at Berkeley, Department of Statistics.

- [Bouvier1994] Bouvier, L., F. R. Fillon and Y. Lespinasse. (1994). Oryzalin as an efficient agent for chromosome doubling of haploid apple shoots in vitro. *Plant Breeding*. 113:343-346.
- [Briggs1997] Briggs, D. and S. M. Walters. 1997. *Plant variation and evolution*, 3rd. ed. Cambridge Univ. Press, Cambridge.
- [Carvalho2006] Carvalho, B. S. et al. (2006), Exploration, normalization, and genotype calls of high-density oligonucleotide SNP array data. *Biostatistics* 8(2), p 485-499.
- [Carvalho2010] Carvalho, P. E. and Stork, D. G. (2001). *Pattern Classification*, 2nd ed. Wiley, New York. Mathematical Reviews (MathSciNet): MR1802993
- [Cleveland79] Cleveland, W. S. (1979). Robust locally weighted regression and smoothing scatterplots. *JASA* 74 829-836.
- [Cleveland88] Cleveland, W. S.; Devlin, S. J. (1988). "Locally-Weighted Regression: An Approach to Regression Analysis by Local Fitting". *Journal of the American Statistical Association* 83 (403): 596–610. doi:10.2307/2289282. JSTOR 2289282. <http://jstor.org/stable/2289282>.
- [CorbachoCorbacho1999] CorbachoCorbacho, J. L., et al. (1999) Differential sequencing with mass spectrometry. *Genetic Analysis: Biomolecular Engineering* 14, 215–219.
- [Ding2004] Ding, C. and Cantor, C.R. (2004) Quantitative analysis of nucleic acids – the last few years of progress. *Journal of Biochemistry and Molecular Biology* 37, 1–10.
- [Ding2004-1] Ding, C., et al. (2004) Simultaneous quantitative and allele-specific expression analysis with real competitive PCR. *BMC Genetics* 5, 8.
- [Downes2004] Downes, K., et al. (2004) SNP allele frequency estimation in DNA pools and variance components analysis. *Biotechniques* 36, 840–845.
- [Dudoit2002-1] Dudoit, S., Fridlyand, J. and Speed, T. (2002). Comparison of discrimination methods for the classification of tumors using gene expression data. *J. Amer. Statist. Assoc.* 97 77–87.
- [Dudoit2002] Dudoit, S. and Fridlyand, J. (2002) ‘A prediction-based resampling method for estimating the number of clusters in a dataset’, *Genome Biology* 3.
- [Edwards2003] Edwards, D. (2003) Non-linear normalization and background correction in onechannel cDNA microarray studies. *Bioinformatics*, 19, 825–833.
- [Ehrich2005] Ehrich, M., et al.(2005) Quantitative high-throughput analysis of DNA methylation patterns by base-specific cleavage and mass spectrometry. *Proceedings of the National Academy of Sciences of the USA* 102, 15785–15790.
- [Gabriel2002] Gabriel, S. B., et al. (2002) The structure of haplotype blocks in the human genome. *Science* 296, 2225–2229.
- [Goldblatt1980] Goldblatt, P. 1980. Polyploidy in angiosperms: monocotyledons, p. 219-239. In: W.H. Lewis (ed.). *Polyploidy*. Plenum Press, New York.
- [Grant1971] Grant, B. 1971. *Plant speciation*. Columbia University Press, New York.
- [Fallah2008] Fallah, S. et al. (2008) Estimating number of clusters based on a general similarity matrix with application to microarray data, *Statistical Applications in Genetics and Molecular Biology*, vol 7, issue 1, article 24.
- [Fujisawa2004] Fujisawa H, Eguchi S, Ushijima M, Miyata S, Miki Y, et al. (2004) Genotyping of single nucleotide polymorphism using model-based clustering. *Bioinformatics (Oxford, England)* 20: 718–26.

- [Harrington2008] Harrington, J. (2008). lga: Tools for linear grouping analysis (LGA). R package version 1.1-1.
- [Hou2007] Hou, W. et al. (2007), Haplotyping a quantitative trait with a high-density map in experimental crosses. *Plos One* 2(8), e732.
- [Huber2002] Huber, W. et al. (2002) Variance stabilization applied to microarray data calibration and to the quantification of differential expression. *Bioinformatics*, 18 (Suppl. 1), S96–S104.
- [Ihaka1996] Ihaka R., Gentleman R. C. (1996), R: a language for data analysis and graphics. *Journal of Computational and Graphical Statistics* 1996 , 5:299-314.
- [Irizarry2003] Irizarry, R. et al. (2003), Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics*, V4, Issue 2, 249-264.
- [Irizarry2003] Kooperberg, C. et al. (2002) Improved background correction for spotted DNA microarrays. *J. Co engineering gene network.ks. Bioinformatics* , 23:i282-288.
- [McGee2006] McGee M., Chen Z. (2006), Parameter Estimation for the Exponential-Normal Convolution Model for Background Correction of Affymetrix GeneChip Data. *Statistical Applications in Genetics and Molecular Biology* 5: Article 24.
- [1] Nadarajah, S. Gupta, A.K. (2006) Some bivariate gamma distributions, *Applied Mathematics Letters*, Volume 19, Issue 8, August 2006, Pages 767-774.
- [RTeam2010] R Development Core Team (2010). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL <http://www.R-project.org>.
- [Ritchie2006] Ritchie, M. I. et al. (2006), A comparison of background correction methods for two-colour microarrays, *Bioinformatics* 23 (20): 2700-2707.
- [Smyth2003] Smyth, G. K., Yang, Y. H., Speed, T. P. (2003). Statistical issues in microarray data analysis. *_Methods in Molecular Biology_* 224, 111-136.
- [Smyth2004] Smyth, G. K. (2004). Linear models and empirical Bayes methods for assessing differential expression in microarray experiments. *_Statistical Applications in Genetics and Molecular Biology_*, Volume 3*, Article 3. <URL: <http://www.bepress.com/sagmb/vol3/iss1/art3>>
- [Smyth2005] Smyth, G. K. (2005). Limma: linear models for microarray data. In: *_Bioinformatics and Computational Biology Solutions using R and Bioconductor_*. R. Gentleman, V. Carey, S. Dudoit, R. Irizarry, W. Huber (eds), Springer, New York, 2005.
- [Tibshirani2001] Tibshirani, R. and Walther, G. and Hastie, T. (2001) ‘Estimating the number of clusters in a data set via the gap statistic’, *J. R. Statist. Soc. B* 63, 411–423.
- [VanAelst2006] Van Aelst, S. et al. (2006) ‘Linear Grouping Using Orthogonal Regression’, *Computational Statistics & Data Analysis* 50, 1287–1312.
- [Ziegler2008] Ziegler, A. ; König, I. R. and Thompson, JR. (2008) Biostatistics aspects of genome-wide association studies. *Biometrical Journal* 50(1), p 8-28.
- [Wang2009] Wang, K. et al. (2009) Estimating the number of clusters via system evolution for cluster analysis of gene expression data, *IEEE Trans. Inf. Tech. in Biomedicine*, vol 13, no. 5.
- [Wolfinger2001] Wolfinger, R. D. et al. (2001), Assessing gene significance from cDNA microarray expression data via mixed models. *Comput Biol.* 8(6), p 625-37.
- [Yang2002a] Yang, Y. H. et al. (2002). Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation. *Nucleic Acids Res.* 30:e15.

- [YangSpeed2002] Yang, Y. H., Speed, T. (2002), Design issues for cDNA microarray experiments. *Nat. Rev. Genet.* 3:579-588, 2002.
- [JohnsonWichern] Johnson, R. and Wichern, D. W. (2002) *Applied Multivariate Statistical Analysis*, 5th ed., Prentice-Hall.
- [Pereira2001] Pereira, C. A. B. and Stern, J. M. (2001). Model Selection: Full Bayesian Approach. *Environmetrics*, 12, 559–568.
- [Madruga2001] M. Madruga, L.G.Esteves, S.Wechsler (2001). On the Bayesianity of Pereira-Stern Tests. *Test*,10,291–299. 117,185–198.
- [Ranney2000] Ranney, T. G. (2000). Polyploidy: From evolution to landscape plant improvement. Proc. 11th Conf. Metropolitan Tree Improvement Alliance.
- [Rowley1993] Rowley, G.D. 1993. Rosaceae: The rose family, p. 141-144. In: V.H. Heywood (ed.). *Flowering plants of the world*. Batsford Pub., London.
- [Kehr1996] Kehr, A. E. 1996. Woody plant polyploidy. *Am. Nurseryman* 183(3):38-47.
- [Roose1976] Roose, M. L. and L. D. Gottlieb. 1976. Genetic and biochemical consequences of polyploidy in *Tragopogon*. *Evolution* 30:818-830.
- [Ruzim] Ruzin, S. E. *Plant microtechnique and microscopy*. Oxford University Press, New York.
- [Sanford, 1983] Sanford, J.C. 1983. Ploidy manipulations, p. 100-123. In: J.N. Moore and J. Janick (eds.). *Methods in fruit breeding*. Purdue Univ. Press, West Lafayette, Ind.
- [Sharma1999] Sharma, A.K. and A. Sharma. 1999. *Plant chromosomes: Analysis, manipulation and engineering*. Harwood Academic Pub., Amsterdam.
- [Solov'eva1990] Solov'eva, L. V. 1990. Number of chloroplasts in guard cells of stomata as an indicator of the ploidy level of apple seedlings. *Cytol. Genet.* 24:1-4.
- [Song et al, 1995] Song, K., P. Lu, K. Tank, and T. C. Osborn. 1995. Rapid genome change in synthetic polyploids of Brassica and its implications for polyploid evolution. *Proceedings of the National Acad. of Sci., USA.* 92: 7719-7723.
- [Soltis1993] Soltis, D. E. and P. S. Soltis. 1993. Molecular data and the dynamic nature of polyploidy. *Critical Rev.Plant Sci.* 12:243-273.
- [Stebbins1994] Stebbins, G. L. 1984. Polyploidy and the distribution of the arctic-alpine flora: new evidence and a new approach. *Botanica Helvetica.* 94:1-13.
- [Taylor1984] Taylor, N. L., et al. 1976. Doubling the chromosome number of *Trifolium* species using nitrous oxide. *Crop Sci.* 16:516-518.
- [van Bogaert1975] van Bogaert, G. 1975. A comparison between colchicine induced tetraploid and diploid cultivars of *Lolium* species. In *Ploidy in Fodder Crops*, Neusch, B. (ed.), Eucarpia Report, Zurich.
- [van Tuy1992] van Tuyl, J. M., B. Meijer, and M. P. van Diën. 1992. The use of oryzalin as an alternative for colchicine in in-vitro chromosome doubling of *Lilium* and *Nerine*. *Acta Hort.* 325:625-629.
- [Voorrips2011] Voorrips RE, Gort G, Vosman B (2011) Genotype calling in tetraploid species from bi-allelic marker data using mixture models. *BMC bioinformatics* 12: 172.
- [9] Serang O, Mollinari M, Garcia AAF (2012) Efficient Exact Maximum a Posteriori Computation for Bayesian SNP Genotyping in Polyploids. *PLoS ONE* 7(2): e30906. doi:10.1371/journal.pone.0030906

Apêndice A

Códigos R desenvolvidos

A.1 graficos.r

```
#####  
#  
#  MODELAGEM E SIMULAÇÃO  
#  
#####  
  
set.seed(2501)  
  
par(family="sans")  
  
# Fig 3.6  
# Modelo de uma urna com capacidade n - sem erro  
for (p in 2:5) {  
  A = 0:p  
  B = p - A  
  png(filename=paste('modelo_uma_urna_ploidia_', p, '.png', sep=''))  
  plot(A, B,  
        pch=19,  
        col = "red4",  
        main = paste("Modelo de uma urna - capacidade", p),  
        sub = expression(italic(w==1)),  
        xlab = expression(italic(P[A])),  
        ylab = expression(italic(P[B])))  
  abline(p, -1, col='black', lty=5)  
  dev.off()  
}  
  
# Fig 3.7  
# Modelo de várias urnas urna com capacidade n - sem erro  
for (p in 2:5) {  
  A = 0:5  
  B = p - A  
  png(filename=paste('modelo_varias_urna_ploidia_', p, '.png', sep=''))  
  plot(cbind(A, B),  
        pch=19,
```

```

        col = "red4",
        main = paste("Modelo de várias urnas - capacidade", p),
        sub = expression(italic(w==1)),
        xlab = expression(italic(P[A])),
        ylab = expression(italic(P[B])),
        xlim = c(0,4*p),
        ylim = c(0,4*p))
for (i in 1:10)
  points(i * cbind(A, B))
abline(p, -1, col='black', lty=5)
dev.off()
}

# Fig 3.8
# Modelo de uma urna com capacidade n - com erro aleatório
N = 100
for (n in 2:5) {
  eA = rnorm((p+1)*N, 0, .15)
  eB = rnorm((p+1)*N, 0, .15)

  A = 0:p
  B = p - A
  X = cbind(eA + A, eB + B)

  png(filename=paste('modelo_uma_urna_erro_', p, '.png', sep=''))
  plot(X,
        pch=19,
        col = "red4",
        main = paste("Modelo de uma urna com erros - capacidade", p),
        sub = expression(italic(w==1)),
        xlab = expression(italic(P[A])),
        ylab = expression(italic(P[B])),
        xlim = c(0,p+1), ylim=c(0, p+1))
  abline(p, -1, col='black', lty=5)
  dev.off()
}

# Modelo de várias urnas com capacidade n - com erro aleatório
N = 100
for (p in 2:5) {
  eA = rnorm((p+1)*N, 0, .1)
  eB = rnorm((p+1)*N, 0, .1)
  C = rgamma((p+1)*N, 4, 1)
  w = 1

  A = w*(0:p)
  B = w*(p - A)
  X = C * cbind(eA + A, eB + B)

  # Fig 3.2
  png(filename=paste('modelo_exemplo', p, '.png', sep=''))
  plot(X,

```

```

    pch=19,
    col = "gray",
    main = "Exemplo",
    xlab = expression(italic(P[A])),
    ylab = expression(italic(P[B])),
    xlim = c(0,max(X)), ylim=c(0, max(X)))
dev.off()

# Fig 3.9
png(filename=paste('modelo_varias_urnas', p, '.png', sep=''))
plot(X,
     pch=19,
     col = "red4",
     main = paste("Modelo várias urnas - capacidade", n),
     xlab = expression(italic(P[A])),
     ylab = expression(italic(P[B])),
     xlim = c(0,max(X)), ylim=c(0, max(X)))
dev.off()

U = X[,1]/rowSums(X)

# Fig 4.2
png(filename=paste('hist_modelo_exemplo', p, '.png', sep=''))
hist(U,
     freq = FALSE,
     col = "steelblue",
     main = paste("Histograma U - urnas com capacidade", p),
     xlab = "U",
     ylab = "densidade",
     xlim = c(0,1),
     nclass=50)
dev.off()

png(filename=paste('densidade_modelo_exemplo_ploidia', p, '.png', sep=''))
z = hist(U, nclass=50)
Dens = approxfun(z$breaks[-1], z$density)
densidade = function(x) ifelse(is.na(Dens(x)), 0, Dens(x))
plot(densidade, 0, 1,
     col="steelblue", lwd=3,
     main = paste("Densidade U - urnas com capacidade", p),
     xlab = "U",
     ylab = "densidade",
     xlim = c(0,1))
front = (0:(p-1) + .5)/p
for (ponto in front) abline(v = ponto, lty="dashed")
dev.off()

if (p == 5) {
  for (j in 2:10) {
    png(filename=paste('densidade_modelo_exemplo_ploidia',
                      p, '-', j, '.png', sep=''))
    plot(densidade, 0, 1,

```

```

        col="steelblue", lwd=3,
        main = paste("Densidade U - urnas com capacidade", p),
        sub = paste("ploidia candidata:", j),
        xlab = "U",
        ylab = "densidade",
        xlim = c(0,1))
    front = (0:(j-1) + .5)/j
    #points(front, densidade(front), pch=19, cex=2)
    points(front, densidade(front), lty=19, cex=2)
    for (ponto in front) abline(v = ponto, lty="dashed")
    dev.off()
  }
}

U = U*p

# Fig 4.3
png(filename=paste('hist_modelo_exemplo_ploidia', p, '.png', sep=''))
hist(U,
      freq = FALSE,
      col = "steelblue",
      main = paste("Histograma U - ploidia", n),
      xlab = expression(U[p]),
      ylab = "densidade",
      xlim = c(0,p),
      nclass=50)
dev.off()
}

#####

plot.gen <-
function (PLOIDIA, SNP, pai, mae, filhos) {

  todos = rbind(pai, mae, filhos)

  lpais = c(rep(1, nrow(pai)), rep(2, nrow(mae)))

  pais = rbind(pai, mae)

  MAX = ceiling(max(todos, na.rm=TRUE))
  XLIM = c(0, MAX)
  YLIM = XLIM

  cores = c(rep("blue", sum(lpais == 1)), rep("red", sum(lpais == 2)))
  PCH = c(rep(15, sum(lpais == 1)), rep(17, sum(lpais == 2)))

```

```

plot(1, type="n",
     xlim = XLIM, ylim = YLIM,
     xlab = expression(italic(P[A])),
     ylab = expression(italic(P[B])),
     main = paste("SNP:", SNP),
     sub = paste("Ploidia:", PLOIDIA))

for (j in seq(.5, PLOIDIA, 2)) {
  Y = c(0, max(0, PLOIDIA-j), max(0, PLOIDIA-j-1), 0)
  X = c(0, j, j+1, 0)
  polygon(2*MAX*X, 2*MAX*Y, col='lightgrey', border='NA')
}
abline(h=0, lty=3)
abline(v=0, lty=3)
abline(0, 1, lty=3)

points(filhos[,1], filhos[,2]) #,
points(pais[,1], pais[,2], col = cores, pch = PCH) #,
grid()
}

loadfiles <- function (n, picos=TRUE) {
  arqpais = paste("SugSNP", n, "_parents", sep= '')
  aux = read.table(arqpais)
  lpais = as.numeric(aux[,1])
  arqfilhos = paste("SugSNP", n, "_progeny", sep= '')
  aux2 = read.table(arqfilhos)

  if (picos) ind = c(2,3)
  else      ind = c(4,5)

  assign("pai", aux[lpais == 1, ind], envir=.GlobalEnv)
  assign("mae", aux[lpais == 2, ind], envir=.GlobalEnv)
  assign("filhos", aux2[, ind], envir=.GlobalEnv)
}

# Fig 4.6
loadfiles(225)
filhos = NULL
png('plot_snp225_ploidia_10.png')
plot.gen(10, 225, pai, mae, filhos)
dev.off()
png('plot_snp225_ploidia_9.png')
plot.gen(9, 225, pai, mae, filhos)
dev.off()

```

A.2 polyploid.r

```
# This library is free software; you can redistribute it and/or
```

```

# modify it under the terms of the GNU Library General Public
# License as published by the Free Software Foundation; either
# version 2 of the License, or (at your option) any later version.
#
# This library is distributed in the hope that it will be useful,
# but WITHOUT ANY WARRANTY; without even the implied warranty of
# MERCHANTABILITY or FITNESS FOR A PARTICULAR PURPOSE. See the
# GNU Library General Public License for more details.
#
# You should have received a copy of the GNU Library General
# Public License along with this library; if not, write to the
# Free Foundation, Inc., 59 Temple Place, Suite 330, Boston,
# MA 02111-1307 USA

#####
# FUNCTION:                DESCRIPTION:
# rGenotype                simulates genotypes
#####

rGenotype <-
  function (n, p = 2, prob = c(0.5, 0.5), ABC=FALSE)
{
  # A function implemented by Silvio Rodrigues de Faria Junior

  # Description:
  # Simulates genotype samples from multiallelic locus with ploidy p

  # Arguments:
  #   n : sample size (integer)
  #   p : ploidy (integer). Remember, humans and animals are diploid.
  #   prob : proportion of alleles (real vector).
  #   ABC : representation with letters instead numbers.

  # Value:
  #   Returns a (n X p) matrix with genotypes, one per line.

  # Note:
  #

  # FUNCTION:

  res = sample.int(length(prob), size = n*p, replace = TRUE, prob = prob)
  if (ABC) res = LETTERS[res]
  if (length(prob) == 2) res = res - 1

  matrix(res, ncol=p)
}

# Examples
if (FALSE) {
# 5 size samples
#####
# Tetraploid sample, same alleles proportion.

```

```

rGenotype(5, 4)
#      [,1] [,2] [,3] [,4]
#[1,]    0    0    0    0
#[2,]    1    0    1    1
#[3,]    0    1    1    1
#[4,]    1    1    0    0
#[5,]    1    1    1    1

# Tetraploid sample, different alleles proportion.
rGenotype(5, 4, prob=c(0.2, 0.8))

# Tetraploid sample with 3 alleles.
rGenotype(5, 4, prob=c(0.2, 0.4, 0.4))

# Tetraploid sample with 3 alleles: A, B, C and D.
rGenotype(10, 4, prob=c(0.2, 0.4, 0.3, .1), ABC = TRUE)

}

#####
# FUNCTION:                DESCRIPTION:
#  geno.prop              calculates genotype proportions of a biallelic locus
#####

geno.prop <-
  function (ploidia, p=0.5)
  {
    # A function implemented by Silvio Rodrigues de Faria Junior

    # Description:
    # Calculates genotype proportions of a biallelic locus.

    # Arguments:
    # ploidia : (integer)
    #       p : proportion of allele A (a real in (0,1) interval).

    # Value:
    # Returns a p+1 vector with genotypic proportions

    # FUNCTION:

    res = dbinom(0:ploidia, ploidia, 1-p)
    f <- function(x) paste(c(rep("A", ploidia - x), rep("B", x)), collapse="")
    names(res) = sapply(0:ploidia, f)
    res
  }

# Examples
if (FALSE) {

```



```

# Tetraploid genotypes, same proportions to both alleles.
geno.prop(4)

# Tetraploid genotypes, 30% alleles A
geno.prop(4, .3)

# Tetraploid genotypes, 70% alleles A
geno.prop(4, .7)

}

```

```

#####
# FUNCTION:                DESCRIPTION:
#   rF1                    generates F1 samples
#####

rF1 <-
  function (n, father, mother)
{
  # A function implemented by Silvio Rodrigues de Faria Junior

  # Description:
  # Calculates genotype proportions of a biallelic locus.

  # Arguments:
  #   n : sample size (integer)
  # father : father genotype (0 and 1 vector)
  # mother : mother genotype (0 and 1 vector)

  # Notes:
  #   father and mother must have the same length.

  # Value:
  #   Returns a (n X p) matrix with offspring genotypes,
  #   where p = length(mother)

  # FUNCTION:

  nfather = length(father)
  nmother = length(mother)
  if (nfather != nmother)
    stop("Father and mother length genotypes cannot differ.")
  ploidia = nmother
  ploid = ploidia
  if ((ploidia %% 2) != 0)
    ploid = ploidia + 1
  fa.gametes = combn(father, ploid/2)
  mo.gametes = combn(mother, ploid/2)
  Ngametes = ncol(fa.gametes)
  fa.choosen = floor(runif(n, 1, Ngametes+1))
  mo.choosen = floor(runif(n, 1, Ngametes+1))
}

```

```

    A = t(rbind(fa.gametes[,fa.choosen], mo.gametes[, mo.choosen]))
    t(apply(A, 1, sample, size=ploidia))
}

```

```

if (FALSE) {

```

```

# triploid homozygote crossing
rF1(10, c(1,1,1), c(0,0,0))

```

```

# tetraploid homozygote crossing
rF1(10, c(1,1,1,1), c(0,0,0,0))

```

```

# tetraploid heterozygote crossing
rF1(10, c(1,1,0,0), c(1,1,0,0))
}

```

```

#####
# FUNCTION:          DESCRIPTION:
# offspring.prop      calculates Mendelian genotypic proportions given
#                    father and mother genotypes in a biallelic locus
#####

```

```

offspring.prop <-

```

```

  function (father, mother)

```

```

{

```

```

  # A function implemented by Silvio Rodrigues de Faria Junior

```

```

  # Description:

```

```

  # Calculates Mendelian genotypic proportions given
  # father and mother genotypes in a biallelic locus

```

```

  # Arguments:

```

```

  # father : father genotype (0 and 1 vector)
  # mother : mother genotype (0 and 1 vector)

```

```

  # Notes:

```

```

  # father and mother must have the same length.

```

```

  # Value:

```

```

  # Returns a vector with the offspring's allele proportions
  # with the counting of allele A in its name.

```

```

  # FUNCTION:

```

```

  nfather = length(father)

```

```

  nmother = length(mother)

```

```

  if (nfather != nmother)

```

```

    stop("Father and mother length genotypes cannot differ.")

```

```

  ploidia = nmother

```

```

  if ((ploidia %% 2) != 0) even.ploidia = FALSE

```

```

    else even.ploidia = TRUE

```

```

fa.freq = table(colSums(combn(father, ploydia %% 2)))
mo.freq = table(colSums(combn(mother, ploydia %% 2)))
fa.gametes = as.numeric(dimnames(fa.freq)[[1]])
mo.gametes = as.numeric(dimnames(mo.freq)[[1]])
if (even.ploydia) {
  frequency = expand.grid(fa.gametes, mo.gametes)
} else {
  aux.gametes = c(0,1)
  if (sum(father) == ploydia || sum(mother) == ploydia)
    aux.gametes = 1
  if (sum(father) == 0 || sum(mother) == 0)
    aux.gametes = 0
  if (sum(father, mother) > sum(fa.gametes, mo.gametes))
    aux.gametes = c(0,1)
  if (sum(father, mother) <= sum(fa.gametes, mo.gametes))
    aux.gametes = 0
  if (sum(father, mother) == ploydia)
    aux.gametes = c(0,1)

  frequency = expand.grid(fa.gametes, mo.gametes, aux.gametes)
}

alleles = rowSums(frequency)
talleles = table(alleles)
talleles/sum(talleles)
}

if (FALSE) {

offspring.prop (c(0,0),
                 c(0,0))
offspring.prop (c(1,0),
                 c(0,0))
offspring.prop (c(1,0),
                 c(1,0))
offspring.prop (c(1,0),
                 c(1,0))
offspring.prop (c(1,1),
                 c(1,1))

}

#####
# FUNCTION:          DESCRIPTION:
# F1.table           generates the table with genotypic proportions given
#                   by a F1 experiment
#####

F1.table <-
  function (ploydia = 3)

```

```

{
  # A function implemented by Silvio Rodrigues de Faria Junior

  # Arguments:
  # ploydia : integer

  # Description:
  # Generates the table with genotypic proportions given
  # by a F1 experiment

  # Value:
  # Returns a matrix with ploydia + 3 columns.
  # First column indicates the number of A alleles from the father
  # Second column indicates the number of A alleles from the mother
  # Other columns indicates the number of A alleles presented by offspring

  # FUNCTION:

  tabl = NULL
  for (i in 0:ploydia) {
    for (j in i:ploydia) {
      father = c(rep(0, ploydia-i), rep(1, i))
      mother = c(rep(0, ploydia-j), rep(1, j))

      off.prop = offspring.prop(father, mother)

      aux = rep(0, ploydia+1)
      inds = as.numeric(names(off.prop)) + 1
      aux[inds] = off.prop
      tabl = rbind(tabl, c(i, j, aux))
    }
  }
  colnames(tabl) = c('father', 'mother', as.character(0:ploydia))
  tabl
}

if (FALSE) {

F1.table(2)

F1.table(3)

F1.table(4)

F1.table(5)

F1 = list()
for (i in 2:25) {
  print(i)
  F1[[i]] = list(ploydia = i, F1.table(i))
  save(F1, file = "F12.table.RData")
}

}

```

```
#####
# FUNCTION:          DESCRIPTION:
# U                  convex combination projection
#
#####

U <- function(x, p) {
  # A function implemented by Silvio Rodrigues de Faria Junior

  # Arguments:
  # x : (vector of length 2 or a matrix with 2 columns) with positive measures
  # p : (integer) ploydia

  # Description:
  # projects the vector measures x onto convex combination [0,p]

  # Value:
  # returns a vector with values in [0,p]

  # FUNCTION:

  if (is.vector(x)) res = p*x[1]/sum(x)
  else {
    if (nrow(x) == 1) res = p*x[1]/sum(x)
    if (ncol(x) == 2) res = p*x[,1]/rowSums(x)
  }
  names(res) = NULL
  res
}

U <- function(x) {
  # A function implemented by Silvio Rodrigues de Faria Junior

  # Arguments:
  # x : (vector of length 2 or a matrix with 2 columns) with positive measures

  # Description:
  # projects the vector measures x onto convex combination [0,p]

  # Value:
  # returns a vector with values in [0,p]

  # FUNCTION:

  if (is.vector(x)) res = x[1]/sum(x)
  else {
    if (nrow(x) == 1) res = x[1]/sum(x)
```

```

        if (ncol(x) == 2) res = x[,1]/rowSums(x)
    }
    names(res) = NULL
    res
}

```

```

if (FALSE) {

    x = cbind(c(0.23, 4.55, 7.27, 3.32, 3.68),
              c(0.33, 0.55, 1.27, 1.32, 0.68))

    U(x)

}

```

```

#####
# FUNCTION:                DESCRIPTION:
# h                        projected data classifier
#
#####

```

```

h <- function(u, p=NA, parents=NULL) {
  # A function implemented by Silvio Rodrigues de Faria Junior

  # Arguments:
  # u : a vector with measures in [0,p] or [0, ceiling(max(u))]
  # p : (integer) ploydia
  # parents : pair of integers in [0,p]

  # Description:
  # classify a vector in

  # Value:
  # returns a vector with integer values in [0,p]

  # FUNCTION:

  #load("F1.table.RData")
  load("F12.table.RData")
  if (!is.null(parents)) {
    parents = sort(parents)
    ind1 = F1[[p]][[2]][, 1] == parents[1]
    ind2 = F1[[p]][[2]][, 2] == parents[2]
    filhos = F1[[p]][[2]][ind1 & ind2, -c(1,2)]
    progenia = as.numeric(names(which(filhos > 0)))
    m1 = min(progenia)
    m2 = max(progenia)
    res = round(u)
    res[which(res < m1)] = m1
  }
}

```

```

        res[which(res > m2)] = m2
    } else
        res = round(u)
    return(res)
}

if (FALSE) {

    x = cbind(c(0.23, 4.55, 7.27, 3.32, 3.68),
              c(0.33, 0.55, 1.27, 1.32, 0.68))

    p = 2
    u = U(x, p)
    h(u)
    parents = c(0,1)
    h(u, p, parents)

    p = 4
    u = U(x, p)
    h(u)
    parents = c(0,1)
    h(u, p, parents)

    p = 6
    u = U(x, p)
    h(u)
    parents = c(0,2)
    h(u, p, parents)

    p = 10
    u = U(x, p)
    h(u)
    parents = c(0,3)
    h(u, p, parents)

}

#####
# FUNCTION:                DESCRIPTION:
# H                        genotyping function
#
#####

H = function(x, p, parents=NULL) {
    # A function implemented by Silvio Rodrigues de Faria Junior

    # Arguments:
    # x : (vector of length 2 or a matrix with 2 columns) with positive measures

```

```

# p : (integer) ploidy
# parents : pair of integers in [0,p]

# Description:
# classify a vector in

# Value:
# returns a vector with integer values in [0,p]

# FUNCTION:

if (!is.null(parents)) hzinho = h(U(x, p), p, parents)
else hzinho = h(U(x, p))
cbind(A = hzinho, B = p-hzinho)
}

```

```

if (FALSE) {

  x = cbind(c(0.23, 4.55, 7.27, 3.32, 3.68),
            c(0.33, 0.55, 1.27, 1.32, 0.68))

  p = 2
  parents = c(0,1)
  H(x, p, parents)

  p = 4
  parents = c(0,1)
  H(x, p, parents)

  p = 6
  parents = c(0,2)
  H(x, p, parents)

  p = 10
  parents = c(0,3)
  H(x, p, parents)

}

```

```

#####
# FUNCTION:          DESCRIPTION:
# most.frequent      most frequent genotype in repeated measures
#
#####

most.frequent <- function(x) as.numeric(names(which.max(table(x))))

if (FALSE) {

  x = c(2, 2, 2, 2, 3, 2, 2, 2, 2, 5)

```



```

    most.frequent(x)

    x = c(3, 3, 3, 3, 3, 3, 2, 2, 2, 5)
    most.frequent(x)
}

#####
# FUNCTION:          DESCRIPTION:
#  misclassified      number of misclassified measures
#
#####

misclassified <-
function(x) abs(length(x) - length(which(x == most.frequent(x))))

if (FALSE) {

    x = c(2, 2, 2, 2, 3, 2, 2, 2, 2, 5)
    misclassified(x)

    x = c(3, 3, 3, 3, 3, 3, 2, 2, 2, 5)
    misclassified(x)

}

#####
# FUNCTION:          DESCRIPTION:
#  distp             distance measures
#
#####

distp <- function (x, y, type = "mse") {

    # A function implemented by Silvio Rodrigues de Faria Junior

    # Arguments:
    # x : vector
    # y : vector
    # type: (character) "mse" for Mean Square Error or
    #          "chi" for Chi-Square Error

    # Description:
    # mean squared error

    # Value:
    #

    # FUNCTION:

```

```

    if (type == "chi") res = c(chi = sum((x - y)**2 / y, na.rm=TRUE))
    if (type == "mse") res = c(mse = sum((x - y)**2, na.rm=TRUE))
    if (type == "mge")
      res = c(mge = abs(1 - exp(sum(log(1+x)-log(1+y), na.rm=TRUE))))
    res
  }

```

```

if (FALSE) {

  x = cbind(c(0.23, 4.55, 7.27, 3.32, 3.68),
            c(0.33, 0.55, 1.27, 1.32, 0.68))

  p = 5
  u = U(x, p)
  hu = h(u)
  distp(u, hu, type="chi")
  distp(u, hu, type="mse")
  distp(u, hu, type="mge")

}

```

```

#####
# FUNCTION:          DESCRIPTION:
# trid                tridiagonal system matrix
#
#####

trid <- function (p.pi, p) {
  # A function implemented by Silvio Rodrigues de Faria Junior

  # Arguments:
  # p.pi : probability of correct classification
  # p: (integer) ploidia

  # Description:
  # generates the tridiagonal matrix to evaluate the probability of correct
  # classification

  # Value:
  # matrix of probabilities (sums 1 per row)

  # FUNCTION:

  res = matrix(0, p+1, p+1)
  diag(res) = p.pi

```

```

    for (i in 1:(p+1)) {
      for (j in 1:(p+1)) {
        if (i == j+1 || i == j-1) res[i,j] = (1 - p.pi)*.5
        if (i == 1 && j == 2 ) res[i,j] = 1 - p.pi
        if (i == p+1 && j == p ) res[i,j] = 1 - p.pi
      }
    }
  }
  res
}

```

```

if (FALSE) {

```

```

  p = 2
  p.pi = 0.9
  trid(p.pi, p)

```

```

  p = 6
  p.pi = 0.85
  trid(p.pi, p)

```

```

}

```

```

#####
# FUNCTION:          DESCRIPTION:
# trid               tridiagonal system matrix
#
#####

```

```

probclass <- function(g, p, p.pi, p.C = NULL) {
  # A function implemented by Silvio Rodrigues de Faria Junior

```

```

  # Arguments:
  # g : vector of genotypes
  # p : ploidy
  # p.pi : probability of correct classification
  # p.C : probability of classification

```

```

  # Description:
  # evaluates the probability of correct classification

```

```

  # Value:
  # vector of probabilities of correct classification

```

```

  # FUNCTION:

```

```

  if (p.pi == 1) p.pi = 0.999

```

```

  if (is.null(p.C)) {
    p.c = table(g) / sum(table(g))
    p.C = rep(0, p+1)
    p.C[as.numeric(names(p.c)) + 1] = p.c
  }

```

```

A = trid(p.pi, p)
b = p.C
fn <- function(x) sum((A %*% x - b) ** 2) + (sum(x) - 1)**2
t0 = c(1, rep(0, p))
lower = rep(0, p+1)
res = optim(par=t0, fn=fn, gr=NULL, method="L-BFGS-B", lower=lower)
p.V = res$par

c(prob = exp(log(p.pi) + log(p.V) - log(p.C)))
}

```

```

if (FALSE) {

  set.seed(2501)
  M = cbind(runif(100), runif(100))
  g = h(U(M, p))
  p = 4
  p.pi = .89
  probclass(g, p, p.pi)

  p = 6
  p.pi = .89
  probclass(g, p, p.pi)

  p = 8
  g = h(U(M, p))
  p.pi = .95
  probclass(g, p, p.pi)

}

```

```

#####
# FUNCTION:          DESCRIPTION:
# fbst              Full Bayesian Significance Test for
#                   Hardy-Weinberg Equilibrium Law in
#                   dialelic polyploid organisms
#####

fbst <- function (p, g, N = 10000, alpha=.10) {
  # A function implemented by Silvio Rodrigues de Faria Junior

  # Arguments:
  # p : ploidy (positive integer)
  # g : vector of genotypes (positive integers)
  # N : (integer) number of sample size for FBST estimation

  # Description:
  # evaluates the evidence against the hypothesis of
  # Hardy-Weinberg equilibrium

```

```

# Value:
# vector of probabilities of correct classification

# FUNCTION:

require(gtools) # for rdirichlet

# genotype counting
x = table(g, useNA="no")
indN = as.numeric(names(x))
n = rep(0, p+1)
n[as.numeric(names(x)) + 1] = x
n = n + 1
# pi_1 as described in Appendix D.1.2, Eq (28)
p1 = sum((0:p)*n)/(p*sum(n))

# loglik under H0
logH0 = sum((0:p)*n*log(p1) + (p:0)*n*log(1-p1))

z = rdirichlet(N, n)
zd = rowSums(n * log(z))
T = length(n)
c(fbst = mean(logH0 < zd), evCr = pchisq(qchisq(1 - alpha, T - 1), T))
}

if (FALSE) {

  set.seed(2501)
  p = 4
  g = round(U(cbind(runif(100), runif(100))),p))
  fbst(p, g)

  p = 8
  g = round(U(cbind(runif(100), runif(100))), p))
  fbst(p, g)
}

plot.gen <-
function (PLOIDIA, SNP, pai, mae, filhos) {

  todos = rbind(pai, mae, filhos)

  lpais = c(rep(1, nrow(pai)), rep(2, nrow(mae)))

  pais = rbind(pai, mae)

  MAX = ceiling(max(todos, na.rm=TRUE))
  XLIM = c(0, MAX)
  YLIM = XLIM

```

```

cores = c(rep("blue", sum(lpais == 1)), rep("red", sum(lpais == 2)))
PCH = c(rep(15, sum(lpais == 1)), rep(17, sum(lpais == 2)))

plot(1, type="n",
     xlim = XLIM, ylim = YLIM,
     xlab = expression(italic(P[A])),
     ylab = expression(italic(P[B])),
     main = paste("SNP:", SNP),
     sub = paste("Ploidia:", PLOIDIA))

for (j in seq(.5, PLOIDIA, 2)) {
  Y = c(0, max(0, PLOIDIA-j), max(0, PLOIDIA-j-1), 0)
  X = c(0, j, j+1, 0)
  polygon(2*MAX*X, 2*MAX*Y, col='lightgrey', border='NA')
}
abline(h=0, lty=3)
abline(v=0, lty=3)
abline(0, 1, lty=3)

points(filhos[,1], filhos[,2]) #,
points(pais[,1], pais[,2], col = cores, pch = PCH) #,
grid()
}

```

A.3 simulation.r

```

#####
# SIMULATIONS
#
# RANDOM SAMPLE
#
# CROSS F1
#
# Implemented by Silvio Rodrigues de Faria Junior

source('polyploid.r')

library(hexbin)

#####
# SIMULAÇÃO
# AMOSTRA ALEATÓRIA
# Ploidias (p): 2, 6 e 12
# Proporções Alélicas (pA): 0.30 e 0.05
# Desvios (eps): 0.75; 0.50; 0.25 e 0.10
#####
# Modelo de soma de erros. Equação (16). Seção 3.4.1.
#####
set.seed(2501)

N = 100000
Rw = rweibull(N, 10, 40)

```

```

w = 1

aas.data = list()
i = 1
for (ploidia in c(2, 6, 12)) {
  for (pA in c(.30, .05)) {
    AA = rGenotype(N, ploidia, c(1-pA, pA))
    a = rowSums(AA)
    b = ploidia - a
    for (eps in c(.75, .50, .25, .1)) {
      EPS = matrix(rnorm(N * ploidia, 0, eps), nrow=N)
      P = abs(cbind(rowSums(AA * (EPS + w)),
                    rowSums((1-AA) * (EPS + w))))
      Pn = P/rowSums(P)
      PAB = Rw * Pn

      arquivo = paste('amostraAleatoria-1-', i, '.png', sep='')
      png(arquivo)
      plot(hexbin(PAB[,1], PAB[,2], xbins=50,
                  xbnds = c(0, 55), ybnds = c(0, 55),
                  xlab = expression(P[A]), ylab = expression(P[B])),
           main = "Amostra Aleatória Simples")
      dev.off()

      simulacoes[[i]] = list(ploidia = ploidia, pA = pA,
                             aas.data[[i]] = list(ploidia = ploidia, pA = pA,
                                                    eps=eps, PAB = cbind(PAB, a, b))

      i = i + 1
    }
  }
}

save(aas.data, file="amostraAleatoria.RData")

#####
# SIMULAÇÃO
# CRUZAMENTOS F1
# Ploidias (p): 10
# Proporções Alélicas (pA): 0.30 e 0.05
# Desvios (eps): 0.75; 0.50; 0.25 e 0.10
#####
# Modelo de soma de erros. Equação (16). Seção 3.4.1.
#####
cruzamentos = list()
# CRUZAMENTO 10A x 7A
cruzamentos[[1]] = list(c(1,1,1,1,1,1,1,1,1,1), c(1,1,1,1,1,1,1,0,0,0))
# CRUZAMENTO 10A x 5A
cruzamentos[[2]] = list(c(1,1,1,1,1,1,1,1,1,1), c(1,1,1,1,1,0,0,0,0,0))
# CRUZAMENTO 10A x 3A
cruzamentos[[3]] = list(c(1,1,1,1,1,1,1,1,1,1), c(1,1,1,0,0,0,0,0,0,0))
# CRUZAMENTO 8A x 3A
cruzamentos[[4]] = list(c(1,1,1,1,1,1,1,1,0,0), c(1,1,1,0,0,0,0,0,0,0))

ploidia = 10

```

```

simulacoes2 = list()
f1.data = list()
i = 1
for (cr in cruzamentos) {
  GF = rF1 (N, cr[[1]], cr[[2]])
  a = rowSums(GF)
  b = ploidia - a
  for (eps in c(.75, .50, .25, .1)) {
    EPS = matrix(rnorm(N * ploidia, 0, eps), nrow=N)
    P = abs(cbind(rowSums(GF * (EPS + w)), rowSums((1-GF) * (EPS + w))))
  )
  Pn = P/rowSums(P)
  PAB = Rw * Pn

  arquivo = paste('cruzamentosF1-1-', i, '.png', sep='')
  png(arquivo)
  plot(hexbin(PAB[,1], PAB[,2], xbins=50,
    xbnds = c(0, 55), ybnds = c(0, 55),
    #xbnds = c(0, max(PAB)), ybnds = c(0, max(PAB)),
    xlab = expression(P[A]), ylab = expression(P[B])),
    main = "Cruzamentos F1")
  dev.off()

  GPai = matrix(rep(cr[[1]], N), nrow=N, byrow=TRUE)
  Pai = abs(cbind(rowSums(GPai * (EPS + w)),
    rowSums((1-GPai) * (EPS + w)))) )
  Pain = Pai/rowSums(Pai)
  PABPai = Rw * Pain

  GMae = matrix(rep(cr[[1]], N), nrow=N, byrow=TRUE)
  Mae = abs(cbind(rowSums(GMae * (EPS + w)),
    rowSums((1-GMae) * (EPS + w)))) )
  Maen = Mae/rowSums(Mae)
  PABMae = Rw * Maen

  #simulacoes2[[i]] = list(cruz = cr, eps=eps, PAB = cbind(PAB, a, b),
  f1.data[[i]] = list(cruz = cr, eps=eps, PAB = cbind(PAB, a, b),
    PABPai = PABPai, PABMae=PABMae)

  i = i + 1
}
}

save(f1.data, file="cruzamentosF1.RData")
#####
#####
#####
# SIMULAÇÃO
# AMOSTRA ALEATÓRIA
# Ploidias (p): 2, 6 e 12
# Proporções Alélicas (pA): 0.30 e 0.05
# Desvios (eps): 0.75; 0.50; 0.25 e 0.10
#####
# Modelo simplificado (variação). Equação (17). Seção 3.4.1.

```



```
#####
set.seed(2501)

N = 100000
Rw = rweibull(N, 10, 40)

#simulacoes = list()
aas.data = list()
i = 1
for (ploidia in c(2, 6, 12)) {
  for (pA in c(.30, .05)) {
    AA = rGenotype(N, ploidia, c(1-pA, pA))
    a = rowSums(AA)
    b = ploidia - a
    for (eps in c(.75, .50, .25, .1)) {
      EPS = rnorm(N, 0, eps)
      P = abs(cbind(a + EPS, b - EPS))
      Pn = P/rowSums(P)
      PAB = Rw * Pn

      arquivo = paste('amostraAleatoria-2-', i, '.png', sep='')
      png(arquivo)
      plot(hexbin(PAB[,1], PAB[,2], xbins=50,
        xbnnds = c(0, 55), ybnnds = c(0, 55),
        #xbnnds = c(0, max(PAB)), ybnnds = c(0, max(PAB)),
        xlab = expression(P[A]), ylab = expression(P[B])),
        main = "Amostra Aleatória Simples")
      dev.off()

      aas.data[[i]] = list(ploidia = ploidia, pA = pA,
        eps=eps, PAB = cbind(PAB, a, b))

      i = i + 1
    }
  }
}

save(aas.data, file="amostraAleatoria.2.RData")

#####
# SIMULAÇÃO
# CRUZAMENTOS F1
# Ploidias (p): 2, 6 e 12
# Proporções Alélicas (pA): 0.30 e 0.05
# Desvios (eps): 0.75; 0.50; 0.25 e 0.10
#####
# Modelo de soma de erros. Equação (16). Seção 3.4.1.
#####
cruzamentos = list()
# CRUZAMENTO 10A x 7A
cruzamentos[[1]] = list(c(1,1,1,1,1,1,1,1,1,1,1), c(1,1,1,1,1,1,1,0,0,0))
# CRUZAMENTO 10A x 5A
cruzamentos[[2]] = list(c(1,1,1,1,1,1,1,1,1,1,1), c(1,1,1,1,1,0,0,0,0,0))
# CRUZAMENTO 10A x 3A
cruzamentos[[3]] = list(c(1,1,1,1,1,1,1,1,1,1,1), c(1,1,1,0,0,0,0,0,0,0))
```

```

# CRUZAMENTO 8A x 3A
cruzamentos[[4]] = list(c(1,1,1,1,1,1,1,1,1,0,0), c(1,1,1,0,0,0,0,0,0,0))

ploidia = 10
#simulacoes2 = list()
f1.data = list()
i = 1
for (cr in cruzamentos) {
  GF = rF1 (N, cr[[1]], cr[[2]])
  a = rowSums(GF)
  b = ploidia - a
  for (eps in c(.75, .50, .25, .1)) {
    EPS = rnorm(N, 0, eps)
    P = abs(cbind(a + EPS, b - EPS))
    Pn = P/rowSums(P)
    PAB = Rw * Pn

    arquivo = paste('cruzamentosF1-2-', i, '.png', sep='')
    png(arquivo)
    plot(hexbin(PAB[,1], PAB[,2], xbins=50,
      xbnnds = c(0, 55), ybnnds = c(0, 55),
      #xbnds = c(0, max(PAB)), ybnnds = c(0, max(PAB)),
      xlab = expression(P[A]), ylab = expression(P[B])),
      main = "Cruzamentos F1")
    dev.off()

    EPS = rnorm(N, 0, eps)
    P = abs(cbind(a + EPS, b - EPS))
    Pn = P/rowSums(P)
    PABPai = Rw * Pn

    EPS = rnorm(N, 0, eps)
    P = abs(cbind(a + EPS, b - EPS))
    Pn = P/rowSums(P)
    PABMae = Rw * Pn

    f1.data[[i]] = list(cruz = cr, eps=eps, PAB = cbind(PAB, a, b),
      PABPai = PABPai, PABMae=PABMae)

    i = i + 1
  }
}

save(f1.data, file="cruzamentosF1.2.RData")

```

A.4 experimentoF1.r

```

source("polyploid.r")

load("cruzamentosF1.2.RData")

loadF1 <- function (i, npais, nfilhos) {

```

```

PAB = f1.data[[i]]$PAB
PABPai = f1.data[[i]]$PABPai
PABMae = f1.data[[i]]$PABMae
N = nrow(PAB)

genPai = sum(f1.data[[i]]$cruz[[1]])
genMae = sum(f1.data[[i]]$cruz[[2]])

indPai = which(PAB[,3] == genPai)
indMae = which(PAB[,3] == genMae)

pai = PABPai[sample.int(N, npais),]
mae = PABMae[sample.int(N, npais),]
filhos = PAB[sample.int(N, nfilhos),]

assign("pai", pai, envir=.GlobalEnv)
assign("mae", mae, envir=.GlobalEnv)
assign("filhos", filhos, envir=.GlobalEnv)

}

```

```

SIMU = 16

```

```

for (i in 16:1) {

  flag = 0

  genPai = sum(f1.data[[i]]$cruz[[1]])
  genMae = sum(f1.data[[i]]$cruz[[2]])

  eps = f1.data[[i]]$eps

  npais = 12
  nfilhos = 180

  for (NSIM in 1:500) {

    loadF1(i, npais, nfilhos)

    Upai = U(pai[,c(1,2)])
    Umae = U(mae[,c(1,2)])
    Ufil = U(filhos[,c(1,2)])

    filtro = 1:length(Ufil)
    if (is.element(NaN, Ufil)) {
      indNaN = which(is.nan(Ufil))
      filtro = filtro[-indNaN]
    }
  }
}

```

```

z = hist(Ufil, nclass=50)
Dens = approxfun(z$breaks[-1], z$density)
densidade = function(x) ifelse(is.na(Dens(x)), 0, Dens(x))

TABELA = NULL
for (p in 2:22) {
  Gpai = h(p * Upai)
  Gmae = h(p * Umae)

  mis.pai = misclassified(p * Upai)
  mis.mae = misclassified(p * Umae)
  Nrep = length(c(pai, mae))

  MF.pai = most.frequent(h(p * Upai))
  MF.mae = most.frequent(h(p * Umae))

  Gfil = h(p * Ufil[filtro], p, c(MF.pai, MF.mae))
  Gfil2 = h(p * Ufil[filtro])

  genotipos = as.numeric(names(table(Gfil)))

  D = as.matrix(dist(cbind(Ufil[filtro], Gfil2/p)))/2
  Dint = 0
  for (g in genotipos)
    Dint = Dint + sum(D[Gfil == g, Gfil == g], na.rm=TRUE)
  Dext = sum(D, na.rm=TRUE) - Dint

  front = (0:(p-1) + .5)/p
  FR = mean(densidade(front)**2)

  chi = distp(Ufil + 1, Gfil/p + 1, type="chi")
  mse = distp(Ufil, Gfil/p, type="mse")
  mge = distp(Ufil, Gfil/p, type="mge")
  fb = fbst(p, Gfil)

  acertosGen = sum(Gfil==filhos[,3], na.rm=TRUE)
  if (p == 10)
    print(acertosGen)

  LINHA = c(p,
    MF.pai, MF.mae,
    misclassified(h(p * Upai)+1),
    misclassified(h(p * Umae)+1),
    sum(abs(Gfil - Gfil2)),
    chi,
    mse,
    mge,
    Dint,
    Dext,
    fb,
    FR,
    acertosGen)

```

```

        TABELA = rbind(TABELA, LINHA)
    }

    rownames(TABELA) = NULL
    colnames(TABELA) = c("p",
                        "G1",      "G2", "ErroG1", "ErroG2", "ErroF",
                        "Quif", "EQMf", "MGEf", "DInterna", "DExterna",
                        "FBST", "evCr", "FR", "acertos")

    message("F1 - i:", i, " npais:", npais, " NSIM:", NSIM)
    tab = cbind(i, npais, NSIM, eps, nfilhos, genPai, genMae, TABELA)

    arquivo = paste("analiseF1", i, ".txt", sep="")

    if (flag == 0) {
        write.table(tab, file=arquivo, append=FALSE, row.names=FALSE, col.names=TRUE,
                    flag = 1)
    } else
        write.table(tab, file=arquivo, append=TRUE, row.names=FALSE, col.names=FALSE)

    }
}

```

A.5 experimentoAAS.r

```

source("polyploid.r")

load("cruzamentosF1.2.RData")

loadF1 <- function (i, npais, nfilhos) {

    PAB = f1.data[[i]]$PAB
    PABPai = f1.data[[i]]$PABPai
    PABMae = f1.data[[i]]$PABMae
    N = nrow(PAB)

    genPai = sum(f1.data[[i]]$cruz[[1]])
    genMae = sum(f1.data[[i]]$cruz[[2]])

    indPai = which(PAB[,3] == genPai)
    indMae = which(PAB[,3] == genMae)

    pai = PABPai[sample.int(N, npais),]
    mae = PABMae[sample.int(N, npais),]
    filhos = PAB[sample.int(N, nfilhos),]

    assign("pai", pai, envir=.GlobalEnv)
    assign("mae", mae, envir=.GlobalEnv)
    assign("filhos", filhos, envir=.GlobalEnv)

}

```

```

SIMU = 16

for (i in 16:1) {

  flag = 0

  genPai = sum(f1.data[[i]]$cruz[[1]])
  genMae = sum(f1.data[[i]]$cruz[[2]])

  eps = f1.data[[i]]$eps

  npais = 12
  nfilhos = 180

  for (NSIM in 1:500) {

    loadF1(i, npais, nfilhos)

    Upai = U(pai[,c(1,2)])
    Umae = U(mae[,c(1,2)])
    Ufil = U(filhos[,c(1,2)])

    filtro = 1:length(Ufil)
    if (is.element(NaN, Ufil)) {
      indNaN = which(is.nan(Ufil))
      filtro = filtro[-indNaN]
    }

    z = hist(Ufil, nclass=50)
    Dens = approxfun(z$breaks[-1], z$density)
    densidade = function(x) ifelse(is.na(Dens(x)), 0, Dens(x))

    TABELA = NULL
    for (p in 2:22) {
      Gpai = h(p * Upai)
      Gmae = h(p * Umae)

      mis.pai = misclassified(p * Upai)
      mis.mae = misclassified(p * Umae)
      Nrep = length(c(pai, mae))

      MF.pai = most.frequent(h(p * Upai))
      MF.mae = most.frequent(h(p * Umae))

      Gfil = h(p * Ufil[filtro], p, c(MF.pai, MF.mae))
      Gfil2 = h(p * Ufil[filtro])

      genotipos = as.numeric(names(table(Gfil)))
    }
  }
}

```

```

D = as.matrix(dist(cbind(Ufil[filtro], Gfil2/p)))/2
Dint = 0
for (g in genotipos)
  Dint = Dint + sum(D[Gfil == g, Gfil == g], na.rm=TRUE)
Dext = sum(D, na.rm=TRUE) - Dint

front = (0:(p-1) + .5)/p
FR = mean(densidade(front)**2)

chi = distp(Ufil + 1, Gfil/p + 1, type="chi")
mse = distp(Ufil,      Gfil/p,      type="mse")
mge = distp(Ufil,      Gfil/p,      type="mge")
fb = fbst(p, Gfil)

acertosGen = sum(Gfil==filhos[,3], na.rm=TRUE)
if (p == 10)
  print(acertosGen)

LINHA = c(p,
  MF.pai, MF.mae,
  misclassified(h(p * Upai)+1),
  misclassified(h(p * Umae)+1),
  sum(abs(Gfil - Gfil2)),
  chi,
  mse,
  mge,
  Dint,
  Dext,
  fb,
  FR,
  acertosGen)

TABELA = rbind(TABELA, LINHA)
}

rownames(TABELA) = NULL
colnames(TABELA) = c("p",
  "G1",      "G2", "ErroG1", "ErroG2", "ErroF",
  "Quif", "EQMf", "MGEf", "DInterna", "DExterna",
  "FBST", "evCr", "FR", "acertos")

message("F1 - i:", i, " npais:", npais, " NSIM:", NSIM)
tab = cbind(i, npais, NSIM, eps, nfilhos, genPai, genMae, TABELA)

arquivo = paste("analiseF1", i, ".txt", sep="")

if (flag == 0) {
  write.table(tab, file=arquivo, append=FALSE, row.names=FALSE, col.names=TRUE)
  flag = 1
} else
  write.table(tab, file=arquivo, append=TRUE, row.names=FALSE, col.names=FALSE)
}

```

```
}
```

A.6 report.r

```
library(xtable)

flag = 0
TABELA = NULL
for (arq in 1:16) {
  arquivo = paste("analiseF1", arq, ".txt", sep="")
  tabela = read.table(arquivo, header=TRUE)

  for (NSIM in 1:500) {
    message("arq:", arq, " NSIM:", NSIM)
    ind = tabela$NSIM == NSIM
    sim = tabela[ind,]
    EQMf = (2:22)[which.min(sim$EQMf)]
    Quif = (2:22)[which.min(sim$Quif)]
    MGEf = (2:22)[which.min(sim$MGEf)]
    FR = (2:22)[which.min(sim$FR)]
    L1 = (2:22)[which.min(sim$FR + sim$EQMf)]
    L2 = (2:22)[which.min(sim$FR + sim$Quif)]
    L3 = (2:22)[which.min(sim$FR + sim$MGEf)]
    aL1 = sim$acertos[L1-1]
    aL2 = sim$acertos[L2-1]
    aL3 = sim$acertos[L3-1]

    genPai = sim$genPai[1]
    genMae = sim$genMae[1]
    nfilhos = sim$nfilhos[1]
    eps = sim$eps[1]

    p = 10

    LINHA = c(arq, NSIM, eps, genPai, genMae, p,
              EQMf, Quif, MGEf, FR, L1, L2, L3, aL1, aL2, aL3)
    print(LINHA)

    TABELA = rbind(TABELA, LINHA)
  }

  if (flag == 0) {
    write.table(TABELA, file="F1sim.txt", append=FALSE,
               row.names=FALSE, col.names=TRUE)
    flag = 1
  } else
    write.table(TABELA, file="F1sim.txt", append=TRUE,
               row.names=FALSE, col.names=FALSE)
}
```



```

colnames(TABELA) = c("arq", "NSIM", "eps",
                     "genPai", "genMae", "p",
                     "EQMf", "Quif", "MGEf", "FRf",
                     "L1", "L2", "L3",
                     "aL1", "aL2", "aL3")
rownames(TABELA) = NULL

TAB = data.frame(TABELA)
save(TABELA, file="F1sim.RData")

load("F1sim.RData")
tab = data.frame(TABELA)
res = NULL
ploidia = 10
for (arq in 1:16) {
  if (arq == 1) {G1 = 10; G2 = 7; eps = 1}
  if (arq == 2) {G1 = 10; G2 = 7; eps = .75}
  if (arq == 3) {G1 = 10; G2 = 7; eps = .50}
  if (arq == 4) {G1 = 10; G2 = 7; eps = .25}

  if (arq == 5) {G1 = 10; G2 = 5; eps = 1}
  if (arq == 6) {G1 = 10; G2 = 5; eps = .75}
  if (arq == 7) {G1 = 10; G2 = 5; eps = .5}
  if (arq == 8) {G1 = 10; G2 = 5; eps = .25}

  if (arq == 9) {G1 = 10; G2 = 3; eps = 1}
  if (arq == 10) {G1 = 10; G2 = 3; eps = .75}
  if (arq == 11) {G1 = 10; G2 = 3; eps = .5}
  if (arq == 12) {G1 = 10; G2 = 3; eps = .25}

  if (arq == 13) {G1 = 8; G2 = 3; eps = 1}
  if (arq == 14) {G1 = 8; G2 = 3; eps = .75}
  if (arq == 15) {G1 = 8; G2 = 3; eps = .5}
  if (arq == 16) {G1 = 8; G2 = 3; eps = .25}

  ta = tab[tab$arq == arq,]

  linha = c(arq = arq, ploidia = ploidia, G1=G1, G2=G2, eps = eps,
            EQMf = sum(ta$EQMf == ploidia, na.rm=TRUE),
            Quif = sum(ta$Quif == ploidia, na.rm=TRUE),
            MGEf = sum(ta$MGEf == ploidia, na.rm=TRUE),
            FR = sum(ta$FR == ploidia, na.rm=TRUE),
            L1 = sum(ta$L1 == ploidia, na.rm=TRUE),
            L2 = sum(ta$L2 == ploidia, na.rm=TRUE),
            L3 = sum(ta$L3 == ploidia, na.rm=TRUE))

  res = rbind(res, linha)
}

write.table(res, file = "reportF1.txt")
save(res, file = "reportF1.RData")

#####

```

```

flag = 0
TABELA = NULL
for (arq in 1:24) {
  arquivo = paste("analiseAA", arq, ".txt", sep="")
  tabela = read.table(arquivo, header=TRUE)

  if (arq == 1) {ploidia = 2; pA = 0.30; eps = .75}
  if (arq == 2) {ploidia = 2; pA = 0.30; eps = .50}
  if (arq == 3) {ploidia = 2; pA = 0.30; eps = .25}
  if (arq == 4) {ploidia = 2; pA = 0.30; eps = .10}
  if (arq == 5) {ploidia = 2; pA = 0.05; eps = .75}
  if (arq == 6) {ploidia = 2; pA = 0.05; eps = .50}
  if (arq == 7) {ploidia = 2; pA = 0.05; eps = .25}
  if (arq == 8) {ploidia = 2; pA = 0.05; eps = .10}

  if (arq == 9) {ploidia = 6; pA = 0.30; eps = .75}
  if (arq == 10) {ploidia = 6; pA = 0.30; eps = .50}
  if (arq == 11) {ploidia = 6; pA = 0.30; eps = .25}
  if (arq == 12) {ploidia = 6; pA = 0.30; eps = .10}
  if (arq == 13) {ploidia = 6; pA = 0.05; eps = .75}
  if (arq == 14) {ploidia = 6; pA = 0.05; eps = .50}
  if (arq == 15) {ploidia = 6; pA = 0.05; eps = .25}
  if (arq == 16) {ploidia = 6; pA = 0.05; eps = .10}

  if (arq == 17) {ploidia = 12; pA = 0.30; eps = .75}
  if (arq == 18) {ploidia = 12; pA = 0.30; eps = .50}
  if (arq == 19) {ploidia = 12; pA = 0.30; eps = .25}
  if (arq == 20) {ploidia = 12; pA = 0.30; eps = .10}
  if (arq == 21) {ploidia = 12; pA = 0.05; eps = .75}
  if (arq == 22) {ploidia = 12; pA = 0.05; eps = .50}
  if (arq == 23) {ploidia = 12; pA = 0.05; eps = .25}
  if (arq == 24) {ploidia = 12; pA = 0.05; eps = .10}

  for (NSIM in 1:500) {
    message("arq:", arq, " NSIM:", NSIM)
    ind = tabela$NSIM == NSIM
    sim = tabela[ind,]
    EQMf = (2:22)[which.min(sim$EQMf)]
    MGEf = (2:22)[which.min(sim$MGEf)]
    Quif = (2:22)[which.min(sim$Quif)]
    FR = (2:22)[which.min(sim$FR)]
    L1 = (2:22)[which.min(sim$FR + sim$EQMf)]
    L2 = (2:22)[which.min(sim$FR + sim$Quif)]
    L3 = (2:22)[which.min(sim$FR + sim$MGEf)]
    aL1 = sim$acertos[L1-1]
    aL2 = sim$acertos[L2-1]
    aL3 = sim$acertos[L3-1]

    LINHA = c(arq, NSIM, eps, ploidia,
              EQMf, Quif, MGEf, FR, L1, L2, L3, aL1, aL2, aL3)
    print(LINHA)
  }
}

```

```

    TABELA = rbind(TABELA, LINHA)

}

if (flag == 0) {
  write.table(TABELA, file="AAsim.txt", append=FALSE,
              row.names=FALSE, col.names=TRUE)
  flag = 1
} else
  write.table(TABELA, file="AAsim.txt", append=TRUE,
              row.names=FALSE, col.names=FALSE)
}

colnames(TABELA) = c("arq", "NSIM", "eps", "p",
                    "EQMf", "Quif", "MGEf", "FR",
                    "L1", "L2", "L3",
                    "aL1", "aL2", "aL3")
rownames(TABELA) = NULL

TAB = data.frame(TABELA)
save(TABELA, file="AAsim.RData")

load("AAsim.RData")

tab = data.frame(TABELA)
res = NULL
for (arq in 1:24) {

  if (arq == 1) {ploidia = 2; pA = 0.30; eps = .75}
  if (arq == 2) {ploidia = 2; pA = 0.30; eps = .50}
  if (arq == 3) {ploidia = 2; pA = 0.30; eps = .25}
  if (arq == 4) {ploidia = 2; pA = 0.30; eps = .10}
  if (arq == 5) {ploidia = 2; pA = 0.05; eps = .75}
  if (arq == 6) {ploidia = 2; pA = 0.05; eps = .50}
  if (arq == 7) {ploidia = 2; pA = 0.05; eps = .25}
  if (arq == 8) {ploidia = 2; pA = 0.05; eps = .10}

  if (arq == 9) {ploidia = 6; pA = 0.30; eps = .75}
  if (arq == 10) {ploidia = 6; pA = 0.30; eps = .50}
  if (arq == 11) {ploidia = 6; pA = 0.30; eps = .25}
  if (arq == 12) {ploidia = 6; pA = 0.30; eps = .10}
  if (arq == 13) {ploidia = 6; pA = 0.05; eps = .75}
  if (arq == 14) {ploidia = 6; pA = 0.05; eps = .50}
  if (arq == 15) {ploidia = 6; pA = 0.05; eps = .25}
  if (arq == 16) {ploidia = 6; pA = 0.05; eps = .10}

  if (arq == 17) {ploidia = 12; pA = 0.30; eps = .75}
  if (arq == 18) {ploidia = 12; pA = 0.30; eps = .50}
  if (arq == 19) {ploidia = 12; pA = 0.30; eps = .25}
  if (arq == 20) {ploidia = 12; pA = 0.30; eps = .10}

```

```

    if (arq == 21) {ploidia = 12;  pA = 0.05; eps = .75}
    if (arq == 22) {ploidia = 12;  pA = 0.05; eps = .50}
    if (arq == 23) {ploidia = 12;  pA = 0.05; eps = .25}
    if (arq == 24) {ploidia = 12;  pA = 0.05; eps = .10}

    ta = tab[tab$arq == arq,]

    linha = c(arq = arq, ploidia = ploidia, pA = pA, eps = eps,
              EQMf = sum(ta$EQMf == ploidia, na.rm=TRUE),
              Quif = sum(ta$Quif == ploidia, na.rm=TRUE),
              MGEf = sum(ta$MGEf == ploidia, na.rm=TRUE),
              FR = sum(ta$FR == ploidia, na.rm=TRUE),
              L1 = sum(ta$L1 == ploidia, na.rm=TRUE),
              L2 = sum(ta$L2 == ploidia, na.rm=TRUE),
              L3 = sum(ta$L3 == ploidia, na.rm=TRUE))

    res = rbind(res, linha)
}

write.table(res, file = "reportAA.txt", sep="\t", dec=',')
save(res, file = "reportAA.RData")

load(file="reportAA.RData")
data = data.frame(res[, -1])
xtab = xtable(data)
digits(xtab) = 0
digits(xtab)[c(3,4)] = 2
align(xtab)[c(1,2,3)] = c("c", "c", "c", "c")
caption(xtab) = c("Simulação AAS")
label(xtab) = "tab:simulacaoAAS"
write(print(xtab), file="simulacaoAAS.tex")

load(file="reportF1.RData")
data = data.frame(res[, -1])
xtab = xtable(data)
digits(xtab) = 0
digits(xtab)[5] = 2
align(xtab)[c(1,2,3)] = c("c", "c", "c", "c")
caption(xtab) = c("Simulação F1")
label(xtab) = "tab:simulacaoF1"
write(print(xtab), file="simulacaoF1.tex")

library(xtable)

load("F1sim.RData")
RES = NULL
ploidia = 10
for (arq in 1:16) {
  if (arq == 1)  {G1 = 10;  G2 = 7; eps = 1}
  if (arq == 2)  {G1 = 10;  G2 = 7; eps = .75}

```

```

if (arq == 3) {G1 = 10; G2 = 7; eps = .50}
if (arq == 4) {G1 = 10; G2 = 7; eps = .25}

if (arq == 5) {G1 = 10; G2 = 5; eps = 1}
if (arq == 6) {G1 = 10; G2 = 5; eps = .75}
if (arq == 7) {G1 = 10; G2 = 5; eps = .5}
if (arq == 8) {G1 = 10; G2 = 5; eps = .25}

if (arq == 9) {G1 = 10; G2 = 3; eps = 1}
if (arq == 10) {G1 = 10; G2 = 3; eps = .75}
if (arq == 11) {G1 = 10; G2 = 3; eps = .5}
if (arq == 12) {G1 = 10; G2 = 3; eps = .25}

if (arq == 13) {G1 = 8; G2 = 3; eps = 1}
if (arq == 14) {G1 = 8; G2 = 3; eps = .75}
if (arq == 15) {G1 = 8; G2 = 3; eps = .5}
if (arq == 16) {G1 = 8; G2 = 3; eps = .25}

ind = TABELA[,1] == arq
tabela = data.frame(TABELA[ind,])
x = tabela$aL1[tabela$L1 == 10]
y = tabela$aL2[tabela$L2 == 10]
z = tabela$aL3[tabela$L3 == 10]
x = x[!is.na(x)]
y = y[!is.na(y)]
z = z[!is.na(z)]
RES = rbind(RES, cbind(arq, G1, G2, eps,
                      rbind(summary(x), summary(y), summary(z))/180))
}

save(RES, file="propacertosF1.RData")

data = data.frame(RES)
xtab = xtable(data)
digits(xtab) = 2
digits(xtab)[c(1,2,3,4)] = 0
align(xtab)[c(1,2,3)] = c("c", "c", "c", "c")
caption(xtab) = c("Proporção de acertos de genotipagem na simulação F1")
label(xtab) = "tab:acertosgenotipagemF1"
write(print(xtab), file="genotipagemF1.tex")

load("AAsim.RData")
RES = NULL
for (arq in 1:24) {

  if (arq == 1) {ploidia = 2; pA = 0.30; eps = .75}
  if (arq == 2) {ploidia = 2; pA = 0.30; eps = .50}
  if (arq == 3) {ploidia = 2; pA = 0.30; eps = .25}
  if (arq == 4) {ploidia = 2; pA = 0.30; eps = .10}
  if (arq == 5) {ploidia = 2; pA = 0.05; eps = .75}
  if (arq == 6) {ploidia = 2; pA = 0.05; eps = .50}

```

```

if (arq == 7) {ploidia = 2;  pA = 0.05; eps = .25}
if (arq == 8) {ploidia = 2;  pA = 0.05; eps = .10}

if (arq == 9) {ploidia = 6;  pA = 0.30; eps = .75}
if (arq == 10) {ploidia = 6;  pA = 0.30; eps = .50}
if (arq == 11) {ploidia = 6;  pA = 0.30; eps = .25}
if (arq == 12) {ploidia = 6;  pA = 0.30; eps = .10}
if (arq == 13) {ploidia = 6;  pA = 0.05; eps = .75}
if (arq == 14) {ploidia = 6;  pA = 0.05; eps = .50}
if (arq == 15) {ploidia = 6;  pA = 0.05; eps = .25}
if (arq == 16) {ploidia = 6;  pA = 0.05; eps = .10}

if (arq == 17) {ploidia = 12;  pA = 0.30; eps = .75}
if (arq == 18) {ploidia = 12;  pA = 0.30; eps = .50}
if (arq == 19) {ploidia = 12;  pA = 0.30; eps = .25}
if (arq == 20) {ploidia = 12;  pA = 0.30; eps = .10}
if (arq == 21) {ploidia = 12;  pA = 0.05; eps = .75}
if (arq == 22) {ploidia = 12;  pA = 0.05; eps = .50}
if (arq == 23) {ploidia = 12;  pA = 0.05; eps = .25}
if (arq == 24) {ploidia = 12;  pA = 0.05; eps = .10}

ind = TABELA[,1] == arq
tabela = data.frame(TABELA[ind,])
x = tabela$aL1[tabela$L1 == ploidia]
y = tabela$aL2[tabela$L2 == ploidia]
z = tabela$aL3[tabela$L3 == ploidia]
x = x[!is.na(x)]
y = y[!is.na(y)]
z = z[!is.na(z)]

RES = rbind(RES, cbind(arq, ploidia, pA, eps,
                        rbind(summary(x), summary(y), summary(z))/180))
}

save(RES, file="propacertosAAS.RData")

data = data.frame(RES)
xtab = xtable(data)
digits(xtab) = 2
digits(xtab)[c(1,2,3)] = 0
align(xtab)[c(1,2,3)] = c("c", "c", "c", "c")
caption(xtab) = c("Proporção de acertos de genotipagem na simulação AAS")
label(xtab) = "tab:acertosgenotipagemAAS"
write(print(xtab), file="genotipagemAAS.tex")

```

A.7 sugarcane2.r

```

#
# Análise dos dados de cana-de-açúcar
#

source("polyploid.r")

```

```

loadfiles <- function (n, picos=TRUE) {
  arqpais = paste("SugSNP", n, "_parents", sep= '')
  aux = read.table(arqpais)
  lpais = as.numeric(aux[,1])
  arqfilhos = paste("SugSNP", n, "_progeny", sep= '')
  aux2 = read.table(arqfilhos)

  if (picos) ind = c(2,3)
  else      ind = c(4,5)

  assign("pai", aux[lpais == 1, ind], envir=.GlobalEnv)
  assign("mae", aux[lpais == 2, ind], envir=.GlobalEnv)
  assign("filhos", aux2[, ind], envir=.GlobalEnv)
}

numeros = c(4, 5, 13, 37, 41,
            45, 48, 50, 60, 65,
            77, 79, 82, 88, 114,
            117, 122, 136, 151, 162,
            186, 201, 204, 225, 235,
            237, 241)

load('F1.table.RData')
PLOIDIA.MAX = length(F1)

for (num in numeros) {

  arqpais = paste("SugSNP", num, "_parents", sep= '')
  aux = read.table(arqpais)
  lpais = as.numeric(aux[,1])
  arqfilhos = paste("SugSNP", num, "_progeny", sep= '')
  aux2 = read.table(arqfilhos)

  png(paste('plot_SugSNP', num, '_gen.png', sep=''), width = 800, height = 600)
  par(mfrow=c(2,2))

  for (picos in c(TRUE, FALSE)) {

    if (picos) ind = c(2,3)
    else      ind = c(4,5)

    pai = aux[lpais == 1, ind]
    mae = aux[lpais == 2, ind]
    filhos = aux2[, ind]

    Upai = U(pai)
    Umae = U(mae)
    Ufil = U(filhos)
  }
}

```

```

filtro = 1:length(Ufil)
if (is.element(NaN, Ufil)) {
  indNaN = which(is.nan(Ufil))
  filtro = filtro[-indNaN]
}

if (picos)
  z = hist(Ufil, nclass=50, main="Picos", density=TRUE,
    ylab="Densidade",
    xlab="U",
    xlim=c(0,1), col="steelblue")
else
  z = hist(Ufil, nclass=50, main="Áreas", density=TRUE,
    ylab="Densidade",
    xlab="U",
    xlim=c(0,1), col="steelblue")
Dens = approxfun(z$breaks[-1], z$density)
densidade = function(x) ifelse(is.na(Dens(x)), 0, Dens(x))

TABELA = NULL
for (p in 2:PLOIDIA.MAX) {
  Gpai = h(p * Upai)
  Gmae = h(p * Umae)

  mis.pai = misclassified(p * Upai)
  mis.mae = misclassified(p * Umae)
  Nrep = length(c(pai, mae))

  MF.pai = most.frequent(h(p * Upai))
  MF.mae = most.frequent(h(p * Umae))

  Gfil = h(p * Ufil[filtro], p, c(MF.pai, MF.mae))

  front = (0:(p-1) + .5)/p
  FR = mean(densidade(front)**2)

  mse = distp(Ufil, Gfil/p, type="mse")
  L1 = mse + FR

  LINHA = c(ploidia = p, G1 = MF.pai, G2 = MF.mae, L1 = L1)
  TABELA = rbind(TABELA, LINHA)
}

TABELA = data.frame(TABELA)
p.hat = (2:PLOIDIA.MAX)[which.min(TABELA$L1.mse)]

MF.pai = most.frequent(h(p.hat * Upai))
MF.mae = most.frequent(h(p.hat * Umae))

A = c(rep(MF.pai, nrow(pai)), rep(MF.mae, nrow(mae)))
B = p.hat - A
aux = cbind(aux, A, B)

A = h(p.hat * Ufil[filtro], p.hat, c(MF.pai, MF.mae))

```



```

B = p.hat - A
x = rep(NA, length(Ufil))
y = rep(NA, length(Ufil))
x[filtro] = A
y[filtro] = B
aux2 = cbind(aux2, x, y)

if (picos)
  plot.gen(p.hat, paste(num," Picos"), pai, mae, filhos)
else
  plot.gen(p.hat, paste(num, " Áreas"), pai, mae, filhos)

}
dev.off()

colnames(aux) = c('id', 'peak.pA', 'peak.pB', 'area.pA', 'area.pB',
                  'A.peak', 'B.peak', 'A.area', 'B.area')

colnames(aux2) = c('id', 'peak.pA', 'peak.pB', 'area.pA', 'area.pB',
                  'A.peak', 'B.peak', 'A.area', 'B.area')

arqpais = paste("SugSNP", num, "_parents.gen", sep= '')
write.table(aux, file=arqpais, row.names=FALSE)
arqfilhos = paste("SugSNP", num, "_progeny.gen", sep= '')
write.table(aux2, file=arqfilhos, row.names=FALSE)

}

```

Apêndice B

Full Bayesian Significance Test (FBST)

Assumindo que o conjunto de dados x , observado em um experimento é proveniente de uma variável aleatória X com um determinado modelo probabilístico paramétrico, o FBST, apresentado por [?], é um teste intuitivo que possui uma interpretação geométrica.

Para maior clareza, introduziremos algumas definições importantes para a construção do teste.

B.1 Algumas definições

Posteriori: Seja $\theta \in \Theta \subset \mathbf{R}^p$, o vetor de parâmetros do modelo que descreve a variável aleatória X e $L(x|\theta)$ a verossimilhança associada aos dados observados x . Pelo Teorema de Bayes, assumindo uma distribuição a priori $p_0(\theta)$ para os parâmetros do modelo de X , a densidade a posteriori $p(\theta)$, é proporcional ao produto da verossimilhança pela priori $p_0(\theta)$.

$$p(\theta) \propto L(x|\theta)p_0(\theta). \quad (\text{B.1})$$

A constante de normalização $m(x)$, que define a identidade dos elementos expressos acima é dada por:

$$m(x) = \int_{\Theta} L(x|\theta)p_0(\theta)d\theta. \quad (\text{B.2})$$

A consideração de informações prévias sobre a distribuição dos parâmetros caracterizada pela escolha da priori p_0 , fica a critério do pesquisador. Uma análise do impacto da escolha de prioris sobre o poder do FBST é descrita em [?], mas neste trabalho não abordaremos esse tema. Em todos os problemas utilizaremos prioris uniformes sobre Θ .

Hipóteses Precisas: Uma *hipótese* é um subconjunto do espaço Θ , que pode ser definido por uma coleção de funções com domínio em Θ e imagem em $\mathbf{R}$, limitadas por igualdades ou desigualdades. Um conjunto de *hipóteses precisas* H formulada sobre o espaço paramétrico Θ , define um conjunto $\Theta_0 \subset \Theta$, de forma que:

$$\Theta_0 = \{\theta \in \Theta : g(\theta) \leq 0 \wedge h(\theta) = 0\}. \quad (\text{B.3})$$

onde g e h são os vetores de restrições sobre Θ e há pelo menos uma restrição de igualdade. Nosso interesse é testar uma hipótese precisa H contra a hipótese alternativa $\bar{H} = \Theta \setminus \Theta_0$.

Credibilidade: Vamos considerar para φ na imagem da densidade a posteriori, o seguinte conjunto:

$$T_{\varphi} = \{\theta \in \Theta : p(\theta) > \varphi\}. \quad (\text{B.4})$$

Denominamos T_{φ} como o *conjunto dos parâmetros tangente à hipótese*, que possuem imagem em p acima do nível φ (i.e., densidade a posteriori), a sua *credibilidade* $\kappa(\varphi)$ é dada por:

$$\kappa(\varphi) = \int_{T_{\varphi}} p(\theta)d\theta = \int_{\Theta} p_{\varphi}(\theta)d\theta. \quad (\text{B.5})$$

com

$$p_\varphi(\theta) = \begin{cases} p(\theta) & \text{se } p(\theta) > \varphi \\ 0 & \text{c.c.} \end{cases}. \quad (\text{B.6})$$

Para o caso em que a posteriori admite um ponto de máximo único, podemos notar que $\kappa(\varphi)$ define uma medida de distância no intervalo $[0; 1]$, entre o ponto de máximo da posteriori e o complementar do conjunto tangente a φ , denotado por $\bar{T}_\varphi$, pois:

1. Se φ^* for o valor máximo de p , $\kappa(\varphi^*) = 0$;
2. Se $\varphi = 0$, então $\bar{T}_\varphi = \emptyset$ e $\kappa(\varphi) = 1$;
3. Se $0 < \varphi_1 < \varphi_2 < \varphi^*$, então $\kappa(\varphi_1) < \kappa(\varphi_2)$.

Da mesma maneira, para posteriors não-normalizadas, $\kappa(\varphi)$, define uma medida de distância, na escala do intervalo $[0, m(x)]$, onde o conjunto $\bar{T}_\varphi$ está mais próximo de φ^* , se $\kappa(\varphi)$ estiver próximo de 0.

Evidência: Considere θ^* como sendo o ponto de máximo da densidade a posteriori sob a hipótese precisa H , e seu respectivo valor na imagem de p , isto é,

$$\theta^* \in \arg \max_{\theta \in \Theta_0} (p(\theta)), \quad p^* = p(\theta^*). \quad (\text{B.7})$$

Definimos as *evidências a favor e contra a hipótese nula H* como:

$$ev(H) = 1 - \kappa(p^*) \quad \text{e} \quad \bar{ev}(H) = \kappa(p^*). \quad (\text{B.8})$$

Assim sendo, $\bar{ev}$ caracteriza uma medida de distância entre o máximo irrestrito da posteriori (φ^*) e o conjunto $\bar{T}_{\theta^*}$. É importante observar que $\Theta_0 \subset \bar{T}_{\theta^*}$, e portanto, se $\hat{\theta} = \arg \max_{\theta \in \Theta} (p(\theta))$, $\bar{ev}$ mede na escala do intervalo $[0, 1]$ a “distância” entre $\hat{\theta}$ e a hipótese H .

B.2 Consistência e Níveis de Significância

Vamos considerar a distribuição acumulada da evidência contra a hipótese, $\bar{V}(c) = Pr(\bar{ev} \leq c)$. Dado θ^0 , o verdadeiro valor do parâmetro, sob condições apropriadas de regularidade, em [?] demonstra-se que, para tamanhos crescentes de amostra, $n \rightarrow \infty$, as seguintes afirmações são verdadeiras:

- Se H é falsa, $\theta^0 \notin \Theta_0$, então $\bar{ev}$ converge em probabilidade para 1, isto é, $\bar{V}(0 < c < 1) \rightarrow 0$. Logo, $\bar{ev}$ é uma estatística consistente;
- Se H é verdadeira, $\theta^0 \in \Theta_0$, então $\bar{ev}$ converge em distribuição para uma $\chi^2(t)$, onde t é o número de graus de liberdade do espaço paramétrico. O nível de confiança $\bar{V}(c)$ (que também denotaremos por α) pode ser aproximado pela função:

$$QQ(t, h, c) = \chi^2(t - h, (\chi^2)^{-1}(t, c)). \quad (\text{B.9})$$

sendo,

- h : número de graus de liberdade do espaço Θ_0 ,
- χ^2 : a densidade chi-quadrado,
- $(\chi^2)^{-1}$: é o quantil da chi-quadrado.

Note que pelo fato de trabalharmos com conjuntos de hipóteses precisas bem definidas, $\dim(\Theta_0) < \dim(\Theta)$, portanto os graus de liberdade $t - h \geq 1$.

A seguir, apresentamos uma tabela com os valores críticos de $\bar{e}\bar{v}$ calculados fixando os níveis de significância de 0,10 e 0,05.

Tabela B.1: Valores Críticos de $\bar{e}\bar{v}$.

$\dim(\Theta)$	$\dim(\Theta_0)$	$c(\alpha = 0,10)$	$c(\alpha = 0,05)$
1	0	0,900	0,950
2	0	0,900	0,950
2	1	0,741	0,853
3	0	0,900	0,950
3	1	0,769	0,887
3	2	0,560	0,720
4	0	0,900	0,950
4	1	0,818	0,901
4	2	0,669	0,800
4	3	0,391	0,572

Note, a medida em que as dimensões de Θ e Θ_0 aumentam, os valores críticos diminuem.

B.3 Procedimento para o teste e FBST Ilustrado

O critério de decisão para a rejeição da hipótese no FBST pode ser baseado apenas no valor da evidência contra a hipótese $\bar{e}\bar{v}$.

Porém, as definições e propriedades da evidência, apresentadas anteriormente, delineiam um procedimento para FBST, que consiste em calcular $\bar{e}\bar{v}(H)$ e confrontá-lo com o valor crítico encontrado invertendo a função QQ para um nível fixado arbitrariamente.

Para isso temos que resolver dois problemas básicos:

1. Encontrar o ponto de máximo θ^* da posteriori p no conjunto H ;
2. Avaliar a integral definida por $\bar{e}\bar{v}$.

O detalhamento destes procedimentos, intuitivamente simples mas que na prática requerem esforços consideráveis.

B.3.1 Teste de Proporção

Vamos mostrar o problema clássico do lançamento de moedas com o FBST. Seja:

- n : o número de lançamentos,
- x : o número de caras,
- y : o número de coroas.

Tomamos o caso $n = 36$, $x = 12$, $y = 24$ e usamos como priori p_0 a densidade uniforme $U[0;1]$. Nesta situação, a estatística utilizada para o teste é o kernel da verossimilhança (o fator que depende dos resultados amostrais): $L(\pi|x, y) = \pi^x(1 - \pi)^y$. A hipótese nula que se deseja testar é $H : \pi = \frac{1}{2}$. A hipótese alternativa neste caso é $A : \pi \neq \frac{1}{2}$.

Figura B.1: FBST para proporção, $(x, y) = (12, 24)$

A hipótese nula $H : \pi = \frac{1}{2}$ é representada pela linha vertical contínua que passa pelo ponto $\pi = 0,5$ nas abscissas. A curva de nível que caracteriza o conjunto tangente à hipótese é a legendada como *Corte* e ela define o conjunto $T = \{\pi | 0,19 < \pi < 0,5\}$. Note que, neste exemplo, a hipótese nula está em uma região de baixa densidade posterior e, assim, $\bar{e}v_H = 95,9\%$. Confrontando com o valor crítico de 95%, encontrado na Tabela ?? para $\dim(\Theta) = 1$ e $\dim(\Theta_0) = 0$, rejeitamos a hipótese H .

B.3.2 Teste de Homogeneidade

O presente exemplo considera duas amostras independentes em que observou-se para cada uma o número de sucessos e fracassos, $(x, y)_i$, onde i é índice da população. O objetivo é testar se as proporções populacionais de sucesso são iguais.

A seguir listamos as diversas etapas do trabalho estatístico.

- **Parâmetros de Interesse:**

π_i : proporção de sucessos na população i , $i = 1, 2$.

- **Densidade Posteriori:** Observamos as seguintes frequências em duas amostras, sendo uma de cada população. $(x, y)_1 = (4, 16)$ e $(x, y)_2 = (10, 10)$

$$f(\pi_1, \pi_2) \propto \pi_1^{x_1} (1 - \pi_1)^{y_1} \pi_2^{x_2} (1 - \pi_2)^{y_2} \quad (\text{B.10})$$

- **Hipótese e Evidência:** $H : \pi_1 = \pi_2$ e $\bar{e}v_H = 98,8\%$

Figura B.2: Visão do plano (π_1, π_2) para o teste de homogeneidade

A Figura ?? ilustra o conjunto tangente (área sombreada da figura), que dá origem ao cálculo da evidência. A hipótese é representada pela linha contínua que corta o quadrado, e o ponto de máximo da hipótese é o ponto representado pelo asterisco. Neste problema a $\dim(\Theta) = 2$ e $\dim(\Theta_0) = 1$, o que nos leva ao valor crítico de 85,3% para $\bar{e}\bar{v}$ com nível de significância de 5%, e portanto, rejeitamos H .

B.3.3 Modelo Normal

Vamos aplicar o FBST ao modelo mais frequentemente utilizado na Estatística, o do conjunto de dados com distribuição normal com média e variância desconhecidas. O presente exemplo foi extraído de [?]. Serão apresentadas para este problema, 3 exemplos de hipóteses compostas, cujo tratamento decorre naturalmente da formulação do FBST. A seguir listamos as diversas etapas do trabalho estatístico.

- **Parâmetros de Interesse:**

- μ : média,
- τ : precisão, definida como o inverso da variância, ou seja, $\tau = \frac{1}{\sigma^2}$.

- **Densidade Posteriori:** A partir de um conjunto de dados observados de um experimento, chegamos ao seguinte kernel para o modelo normal $N(\mu, \tau)$:

$$f(\mu, \tau) \propto \tau^{6,5} \exp \left\{ -\frac{\tau [4 + 11(\mu - 0,9)^2]}{2} \right\} \quad (\text{B.11})$$

- **Hipóteses e Evidências:**

Tabela B.2: Evidências contra hipóteses do modelo

H_0 :	$\mu = 1, 1$	$\bar{ev}_0 = 0, 49$
H_1 :	$\tau = 2, 5$	$\bar{ev}_1 = 0, 19$
H_2 :	$cv = \sqrt{\mu^2 \tau} = \frac{1}{2}$	$\bar{ev}_2 = 0, 31$
H_3 :	$\mu = 1, 1$ e $\tau = 2, 5$	$\bar{ev}_3 = 0, 53$
H_4 :	$\mu = 1, 1$ e $cv = \frac{1}{2}$	$\bar{ev}_4 = 0, 52$
H_5 :	$\tau = 2, 5$ e $cv = \frac{1}{2}$	$\bar{ev}_5 = 0, 87$

Neste exemplo, as hipóteses H_0 , H_1 e H_2 possuem dimensão 1 no espaço paramétrico. Fixando o nível de significância de 5% e comparando com o valor crítico de 0,853 dado pela Tabela ??, não as rejeitamos.

As hipóteses H_3 , H_4 e H_5 são conjuntos pontuais e possuem dimensão 0 no espaço paramétrico. Neste caso também não as rejeitamos para o nível de significância de 0,05. O valor crítico de $\bar{ev}$ neste caso, é de 0,950.

Figura B.3: Densidade posteriori de $f(\mu, \tau)$:
Curvas das Hipóteses 0, 1 e 2

A Figura ?? apresenta a densidade a posteriori de (μ, τ) e as curvas correspondentes ao espaço paramétrico sob as hipóteses: $H_0 : \mu = 1, 1$ (linha tracejada), $H_1 : \tau = 2, 5$ (linha contínua clara) e $H_2 : cv = \sqrt{\mu^2 \tau} = \frac{1}{2}$ (linha contínua escura).

 Figura B.4: Cortando a posteriori para definir o conjunto T_0

Na Figura ??, observamos o plano tangente à hipótese $H_0 : \mu = 1,1$; definido pela curva de nível $f(\mu, \tau) = f(\theta_{H_0}^*)$. A superfície acima do plano corresponde à imagem de f no conjunto $T_{\Theta_0} \subset \Theta = \{(\mu, \tau) : \mu \in \mathbf{R}, \tau \in \mathbf{R}_*^+\}$

 Figura B.5: Cortando a densidade posteriori para definir os conjuntos T_0 e T_1

Figura B.6: Cortando a densidade posteriori para definir os conjuntos T_0 e T_1 e T_2

A Figura ?? representa o espaço paramétrico Θ e as curvas de nível tangentes às hipóteses apresentadas. Os pontos θ^* de cada hipótese estão listados na Tabela ??.

Figura B.7: Curvas de Nível:
Conjuntos Tangentes T_0 , T_1 , T_2 , T_3 , T_4 , e T_5

Tabela B.3: Pontos de máxima posteriori sob as hipóteses do modelo

H_0 :	$\mu = 1, 1$	$\theta_{H_0}^* = (1, 1; 2, 92)$
H_1 :	$\tau = 2, 5$	$\theta_{H_1}^* = (0, 9; 2, 5)$
H_2 :	$cv = \sqrt{\mu^2 \tau} = \frac{1}{2}$	$\theta_{H_2}^* = (0, 98; 4, 12)$
H_3 :	$\mu = 1, 1$ e $\tau = 2, 5$	$\theta_{H_3}^* = (1, 1; 2, 5)$
H_4 :	$\mu = 1, 1$ e $cv = \frac{1}{2}$	$\theta_{H_4}^* = (1, 1; 3, 30)$
H_5 :	$\tau = 2, 5$ e $cv = \frac{1}{2}$	$\theta_{H_5}^* = (1, 26; 2, 5)$

Apêndice C

O Equilíbrio de Hardy-Weinberg

O problema do equilíbrio de Hardy-Weinberg é clássico nos estudos de genética populacional. Grosso modo, em uma população em equilíbrio, as frequências genotípicas populacionais podem ser calculadas a partir das frequências alélicas, e vice-versa. Uma extensa bibliografia trata do teste de equilíbrio de Hardy-Weinberg é, dentre os quais citamos [?], bem como a solução originalmente proposta pelo FBST, [?].

Antes de tratarmos sobre o *Equilíbrio de Hardy-Weinberg*, a *Evolução Darwiniana* e aplicações do FBST nessa temática, é importante discutirmos sobre alguns fundamentos e conceitos de genética. Na seção seguinte, veremos o *Equilíbrio de Hardy-Weinberg*, uma propriedade importante que relaciona as características genéticas entre gerações de uma população. Após isso formalizaremos o teste de equilíbrio com o FBST. Em seguida trataremos um pouco sobre fatores da dinâmica evolutiva que influenciam o comportamento dos perfis genéticos de uma população ao longo do tempo e sua relação com o *Equilíbrio de Hardy-Weinberg*.

Genética populacional é o estudo das distribuições de frequências dos *alelos* e suas mudanças sob a influência de forças evolutivas tais como: seleção natural, mutação, migração e *deriva genética*. Ela também considera subdivisões populacionais e a estrutura de uma população no espaço.

Gene alelo é qualquer uma das possíveis sequências de DNA em um determinado *locus* (posição cromossômica). Em seus estudos, Gregor Mendel, concluiu que na variedade de ervilhas que cultivava, havia dois alelos para a determinação da cor das flores produzidas por elas, sendo um deles chamado de *dominante* (por forçar a produção de flores de uma cor com sua simples presença em um dos alelos) e o outro de *recessivo* (por determinar a aparição de outra cor apenas quando os dois alelos fossem recessivos simultaneamente), este é o caso mais simples em espécies *diplóides* (espécies cujas características genéticas são determinadas por *pares* de cromossomos). Existem locus que podem admitir dezenas de alelos. Um locus famoso que exemplifica o caso de um gene de 3 alelos é o do que caracteriza a classificação do sangue humano no sistema ABO.

C.1 Equilíbrio de Hardy-Weinberg

Em espécies diplóides, no caso em que um locus possui dois alelos A e a , com frequências populacionais p e q , respectivamente, com $p + q = 1$, se a população não estiver sofrendo ação de alguma das forças evolutivas sobre esse locus, as frequências genotípicas dos pares de alelos AA , Aa e aa em uma nova geração de indivíduos podem ser representadas pela Tabela C.1:

	$A (p)$	$a (q)$
$A (p)$	$AA (p^2)$	$Aa (pq)$
$a (q)$	$Aa (pq)$	$aa (q^2)$

Tabela C.1: Frequências genotípicas esperadas no Equilíbrio de Hardy-Weinberg

A Tabela C.1 indica que esperamos observar na próxima geração as frequências:

$$f(AA) = p^2, \quad f(aa) = q^2, \quad f(Aa) = 2pq,$$

tais que $p^2 + 2pq + q^2 = 1$.

Sob as hipóteses de:

- cruzamentos aleatórios entre os indivíduos da população;
- número de indivíduos suficientemente grande para evitar grandes flutuações aleatórias entre as frequências de duas gerações;
- ausência de influência das forças evolutivas como, seleção natural, mutações e migrações.

consideramos que o locus com frequências alélicas populacionais descritas acima está sob o *Equilíbrio de Hardy-Weinberg*.

C.1.1 Caso Multialélico Diplóide

No caso em que o locus analisado possui mais de 2 alelos em seres diplóides, obtemos as equações de equilíbrio a partir da expansão multinomial de grau 2 das frequências alélicas populacionais. Tomando um locus com k alelos, os fatores da expansão de $(p_1 + \dots + p_k)^2$ são:

$$\left(\underbrace{p_1}_{f(A_1)} + \dots + \underbrace{p_k}_{f(A_k)} \right)^2 = \underbrace{p_1^2}_{f(A_1A_1)} + \dots + \underbrace{p_k^2}_{f(A_kA_k)} + \sum_{i \neq j} \underbrace{2p_i p_j}_{f(A_iA_j)} \quad (C.1)$$

onde os $p_i = f(A_i)$, $i = 1 \dots k$ são as frequências alélicas populacionais para o gene A_i e $f(A_iA_j)$ as frequências dos genes A_iA_j .

Uma vez em equilíbrio, as relações entre as frequências genotípicas e as frequências alélicas devem satisfazer:

$$f(A_iA_j) = f(A_i)f(A_j) \quad \forall i, j \in \{1, 2, \dots, k\}. \quad (C.2)$$

A formalização do FBST para o equilíbrio entre as frequências genotípicas e alélicas na geração de uma população para o caso de organismos *diploides*, pode ser feita da seguinte forma:

- **Posteriori:** Denotemos por $\mathbf{n}$ o vetor de contagem dos genótipos observados em uma amostra:

$$\mathbf{n} = \begin{pmatrix} n_{A_1A_1} & 0 & \dots & 0 \\ n_{A_2A_1} & n_{A_2A_2} & \ddots & \vdots \\ \vdots & & \ddots & 0 \\ n_{A_kA_1} & \dots & \dots & n_{A_kA_k} \end{pmatrix} \begin{pmatrix} 1 \\ \vdots \\ \vdots \\ 1 \end{pmatrix}. \quad (C.3)$$

Admitindo a distribuição a priori uniforme sobre o espaço paramétrico, a distribuição a posteriori para as frequências genotípicas $\pi_{ij} = f(A_iA_j)$ é dada por:

$$p(\boldsymbol{\pi}) \propto \exp \{ \mathbf{n} \cdot \log(\boldsymbol{\pi}) \} = \prod_{i, j \in \{1, \dots, k\}} \pi_{ij}^{n_{A_iA_j}} \quad (C.4)$$

onde $\boldsymbol{\pi}$, é o vetor das frequências genotípicas π_{ij} .

- **Hipótese:** Denotando por π_i , $i = 1, \dots, k$, as frequências alélicas populacionais, nossa hipótese é composta pelas equações de equilíbrio:

$$\pi_{ij} = \pi_i \pi_j \quad \forall i, j \in \{1, 2, \dots, k\}, \quad (C.5)$$

e pela restrição do espaço paramétrico:

$$\sum_{i=1}^k \pi_i = 1. \quad (C.6)$$

Sob H , a posteriori neste caso fica:

$$p(\boldsymbol{\pi}|H) \propto \prod_{\substack{j \in \{1, \dots, k\} \\ i \in \{j, \dots, k\}}} (\pi_i \pi_j)^{n_{A_i A_j}} \quad (\text{C.7})$$

Otimização Para a otimização sob a hipótese, podemos utilizar o método analítico apresentado a seguir. Note-se que a Equação C.7 pode ser reescrita separando-se os genótipos homozigotos (alelos iguais) dos heterozigotos (alelos diferentes):

$$p(\boldsymbol{\pi}|H) \propto f(\boldsymbol{\pi}|H) = \prod_{j \in \{1, \dots, k\}} \pi_j^{2n_{A_j A_j}} \prod_{\substack{j \in \{1, \dots, k-1\} \\ i \in \{j+1, \dots, k\}}} (2\pi_i \pi_j)^{n_{A_i A_j}}$$

Maximizar $f(\boldsymbol{\pi}|H)$ equivale a maximizar seu logaritmo:

$$\begin{aligned} \log(f(\boldsymbol{\pi}|H)) &= l(\boldsymbol{\pi}) = \sum_{j \in \{1, \dots, k\}} 2n_{A_j A_j} \log(\pi_j) + \sum_{\substack{j \in \{1, \dots, k-1\} \\ i \in \{j+1, \dots, k\}}} n_{A_i A_j} \log(2\pi_i \pi_j) \\ &= \sum_{j \in \{1, \dots, k\}} 2n_{A_j A_j} \log(\pi_j) + \sum_{\substack{j \in \{1, \dots, k-1\} \\ i \in \{j+1, \dots, k\}}} n_{A_i A_j} (\log(2) + \log(\pi_i) + \log(\pi_j)) \\ &= h \log(2) + \sum_{j \in \{1, \dots, k\}} n_j \log(\pi_j), \end{aligned}$$

tal que,

$$\begin{aligned} h &= \sum_{\substack{j \in \{1, \dots, k-1\} \\ i \in \{j+1, \dots, k\}}} n_{A_i A_j}, \\ n_j &= \sum_{i \in \{1, \dots, j\}} n_{A_j A_i} + \sum_{i \in \{j, \dots, k\}} n_{A_i A_j}. \end{aligned}$$

Pela restrição C.6, temos:

$$l(\boldsymbol{\pi}) = h \log(2) + \sum_{j \in \{1, \dots, k-1\}} n_j \log(\pi_j) + n_k \log \left(1 - \sum_{j \in \{1, \dots, k-1\}} \pi_j \right).$$

As componentes do gradiente de $l(\boldsymbol{\pi})$ são dadas por:

$$\frac{\partial l}{\partial \pi_i}(\boldsymbol{\pi}) = \frac{n_i}{\pi_i} - \frac{n_k}{\left(1 - \sum_{j=1}^{k-1} \pi_j\right)}$$

Logo, o ponto ótimo sob a hipótese H é o vetor $\boldsymbol{\pi}$ que satisfaz:

$$A \cdot \boldsymbol{\pi} = b; \quad A = \begin{pmatrix} n_1 + n_k & n_1 & \dots & n_1 \\ n_2 & n_2 + n_k & & n_2 \\ \vdots & & \ddots & \vdots \\ n_{k-1} & n_{k-1} & \dots & n_{k-1} + n_k \end{pmatrix}; \quad b = \begin{pmatrix} n_1 \\ \vdots \\ \vdots \\ n_{k-1} \end{pmatrix}.$$

Integração Neste problema, devemos utilizar o Método da Variável Indicadora descrito no capítulo anterior. A amostragem da distribuição Dirichlet definida em (C.4) é simples de ser obtida e podemos aplicar a transformação do logaritmo para calcular $e\bar{v}$.

C.1.2 Caso Dialélico Poliplóide

No caso em que o locus analisado possui 2 alelos em seres poliplóides (de ploidia k), obtemos as equações de equilíbrio a partir da expansão binomial de grau k das frequências alélicas populacionais:

$$\underbrace{\left(\underbrace{p_1}_{f(A)} + \underbrace{p_2}_{f(B)} \right)^k}_{f(G_i)} = \sum_{i=0}^k \underbrace{\binom{k}{i} p_1^i p_2^{k-i}}_{f(G_i)} \quad (\text{C.8})$$

onde:

- $p_1 = f(A)$, é a proporção populacional do alelo A ;
- $p_2 = f(B)$, é a proporção populacional do alelo B ;
- $\binom{k}{i} p_1^i p_2^{k-i} = f(G_i)$, é a proporção populacional do genótipo G_i , com i sendo o número de alelos A presente no genótipo.

onde os $p_i = f(A_i)$, $i = 1, 2$ são as frequências alélicas populacionais para o alelo A_i e $f(A_i A_j)$ as frequências dos possíveis genótipos.

Uma vez em equilíbrio, as relações entre as frequências genotípicas e as frequências alélicas devem satisfazer:

$$f(G_i) = f(A)^i f(B)^{k-i} \quad \forall i, j \in \{1, 2\}. \quad (\text{C.9})$$

A formalização do FBST para o equilíbrio entre as frequências genotípicas e alélicas na geração de uma população para o caso de organismos com nível de ploidia k , pode ser feita da seguinte forma:

- **Posteriori:** Denotemos por $\mathbf{n}$ o vetor de contagem dos genótipos observados em uma amostra:

$$\mathbf{n} = \begin{pmatrix} n_{G_{0,k}} \\ n_{G_{1,k-1}} \\ \vdots \\ n_{G_{k,0}} \end{pmatrix} = \begin{pmatrix} n_0 \\ n_1 \\ \vdots \\ n_k \end{pmatrix}. \quad (\text{C.10})$$

Admitindo a distribuição a priori uniforme sobre o espaço paramétrico, a distribuição a posteriori para as frequências genotípicas $\pi_{ij} = f(A_i A_j)$ é dada por:

$$p(\boldsymbol{\pi}) \propto \exp \{ \mathbf{n} \cdot \log(\boldsymbol{\pi}) \} = \prod_{i=0}^k (\pi_{i,k-i})^{n_i} \quad (\text{C.11})$$

onde $\boldsymbol{\pi}$, é o vetor das frequências genotípicas π_{ij} .

- **Hipótese:** Denotando por π_i , $i = 1, 2$, as frequências alélicas populacionais, nossa hipótese é composta pelas equações de equilíbrio:

$$\pi_{i,k-i} = \pi_1^i \pi_2^{k-i} \quad \forall i \in \{0, \dots, k\}, \quad (\text{C.12})$$

e pela restrição do espaço paramétrico:

$$\pi_1 + \pi_2 = 1. \quad (\text{C.13})$$

Sob H , a posteriori neste caso fica:

$$p(\boldsymbol{\pi}|H) \propto \prod_{i=0}^k (\pi_1^i \pi_2^{k-i})^{n_i} \quad (\text{C.14})$$

$$\propto \exp \left\{ \sum_{i=0}^k (i \cdot n_i \log \pi_1 + (k-i) \cdot n_i \log \pi_2) \right\} \quad (\text{C.15})$$

$$\propto \exp \left\{ \sum_{i=0}^k (i \cdot n_i \log \pi_1 + (k-i) \cdot n_i \log(1 - \pi_1)) \right\} \quad (\text{C.16})$$

Otimização Maximizar $f(\boldsymbol{\pi}|H)$ equivale a maximizar seu logaritmo:

$$\begin{aligned}
 \log(f(\boldsymbol{\pi}|H)) &= l(\boldsymbol{\pi}) = \sum_{i=0}^k (i.n_i \log \pi_1 + (k-i)n_i \log(1-\pi_1)) \\
 \frac{dl}{d\pi_1} &= \frac{1}{\pi_1} \sum_{i=0}^k i.n_i - \frac{1}{1-\pi_1} \sum_{i=0}^k (k-i).n_i = 0 \\
 \Rightarrow &\left(\sum_{i=0}^k i.n_i \right) (1-\pi_1) = \pi_1 \left(\sum_{i=0}^k (k-i).n_i \right) \\
 \Rightarrow &\pi_1 = \frac{\left(\sum_{i=0}^k i.n_i \right)}{\left[\sum_{i=0}^k i.n_i + \sum_{i=0}^k (k-i).n_i \right]} \\
 &\pi_1 = \frac{\sum_{i=0}^k i.n_i}{k \sum_{i=0}^k n_i} = \frac{1}{kN} \sum_{i=0}^k i.n_i
 \end{aligned} \tag{C.17}$$

Portanto a proporção de alelos π , que maximiza $f(\boldsymbol{\pi}|H)$ é dada por:

$$\boldsymbol{\pi} = \begin{pmatrix} \pi_1 \\ \pi_2 \end{pmatrix} = \begin{pmatrix} \frac{1}{kN} \sum_{i=0}^k i.n_i \\ 1 - \frac{1}{kN} \sum_{i=0}^k i.n_i \end{pmatrix}$$

com

$$N = \sum_{i=0}^k n_i$$

C.1.3 Caso Multialélico Poliplóide

A generalização das equações acima para o caso de n -*ploidia* em um locus de k alelos, é dada por:

$$(p_1 + \dots + p_k)^n = \sum_{\substack{l_1, \dots, l_k: l_1 + \dots + l_k = n, \\ l_r \in \{1, 2, \dots, n\}, r \in \{1, 2, \dots, k\}}} \binom{n}{l_1, \dots, l_k} p_1^{l_1} \dots p_k^{l_k}, \tag{C.18}$$

$$f(A_{i_1}, \dots, A_{i_k}) = f(A_{i_1}) \dots f(A_{i_k}) \quad \forall i_r \in \{1, \dots, n\}, r \in \{1, \dots, k\}$$

Apêndice D

Plataforma Affymetrix

Os experimentos com SNPs utilizam a mesma tecnologia envolvida nos experimentos com microarrays de 2 canais. De cada indivíduo o DNA extraído é adicionado a uma lâmina de vidro ou chip onde estão sintetizadas milhares de sondas cujo alvo são os alelos de particulares SNPs de posição conhecida no genoma (em pares de bases ou bp). A plataforma de SNPs da Affymetrix 6.0, principalmente usada para estudos com humanos, é composta por chips que permitem avaliar o genótipo do indivíduo em 1 milhão de SNPs de uma única vez (Figura D.2).

Figura D.2: GeneChip da Affymetrix

Sob esta tecnologia (Figura D.1) a leitura dos genótipos de cada SNP não é direta. Por meio de um processador de imagem são coletadas, de cada sonda no chip de cada indivíduo, leituras de intensidade de luz que correspondem às quantidades dos alelos, A e B por exemplo, que compõem o loco do SNP. Para o caso diplóide, de maneira geral, se há um excesso na quantidade de alelos A relativamente ao alelo B, o genótipo homozigoto AA deve ser indicado; se as quantidades dos alelos são semelhantes, o genótipo heterozigoto AB deve ser indicado; finalmente, se há um excesso na quantidade de alelos B relativamente ao alelo A, o genótipo homozigoto BB deve ser indicado.

Figura D.1: Processo de obtenção da imagem dos chips

Figura D.3: Fenômeno de hibridização das moléculas no chip

Figura D.4: Fenômeno de hibridização das moléculas no chip

Figura D.5: Fenômeno de hibridização das moléculas no chip

Figura D.6: Exemplo de dados de SNPs para o caso diplóide, o genótipo heterozigoto AB forma um agrupamento em torno da reta que demarca níveis de expressão iguais em cada alelo.

Tal categorização das quantidades dos alelos é feita por meio de algoritmos de agrupamento. O processamento dos chips, em geral, é feito em grupos dispostos em uma placa. No caso da plataforma de SNPs da Affymetrics 6.0 por exemplo, o processador tem capacidade para fazer a leitura de 96 chips por vez, os quais ficam dispostos em uma placa de 12x8 lugares. É esperado que ocorram variações indesejáveis nas leituras entre sondas dentro da placa por conta da posição ou devido ao tamanho do fragmento/sonda, bem como variações de placa para placa devido ao desgaste do aparelho, entre outras causas. Deste modo, os seguintes passos são típicos destes estudos:

- **Delineamento experimental:** decisões referentes ao tipo de planejamento experimental a ser adotado na coleta das unidades amostrais (indivíduos não relacionados, pedigrees/famílias, F2, retrocruzamento, etc.), à plataforma de marcadores moleculares (quais SNPs, quais sondas etc.) e, muitas vezes, à aleatorização dos chips nas placas de processamento.
- **Controle de Qualidade:** as quantificações dos alelos A e B realizadas de cada indivíduo em cada sonda são analisadas podendo ocorrer a exclusão de SNPs de baixa qualidade e exclusão de indivíduos com DNA mal extraído ou com leituras de baixa confiabilidade. Para o conjunto de indivíduos, os alelos de mais baixa ocorrência em cada SNP são também identificados. Este passo é bastante similar para organismos diplóides ou poliplóides.
- **Pré-processamento:** as leituras alélicas são corrigidas para remoção do ruído de *background* (leitura de um borrão em cada sonda), normalizadas para remoção dos efeitos de variáveis intervenientes (posição no chip, tamanho do fragmento), e, em alguns casos, há necessidade de redução dos dados de cada SNP (quando mais de uma sonda é usada para um mesmo SNP alvo).
- **Genotipagem:** com os dados pré-processados, algoritmos de classificação e agrupamento são utilizados para atribuir os genótipos aos elementos da amostra, dependendo do organismo em questão, a existência de bases de dados com elementos genotipados “in vitro” facilita o desenvolvimento de técnicas supervisionadas.
- **Análise dos efeitos dos SNPs:** mapeamento genético propriamente dito. Ziegler et al. (2008)[Ziegler2008] discutem diferentes métodos para se fazer inferências sobre os efeitos de SNPs em estudos caso-controle em populações humanas. Aultchenko et al. (2007)[Aultchenko2007] apresentam uma estratégia em multi-estágios para análise de dados de SNPs em dados de famílias/pedigrees. Hou et al. (2007)[Hou2007] propõe uma estratégia de análise de dados de SNPs por haplótipos para populações F2 e indivíduos diplóides.

Apêndice E

Pré-processamento

As leituras alélicas são corrigidas para remoção do *background*. A remoção do background pode ser feita via correções simples pela mediana ou por meio de modelos que assumem distribuições probabilísticas específicas para as leituras do *background* e do *foreground*. Como as leituras alélicas são bastante assimétricas, as análises são feitas na variável log(base 2)-transformada. Os métodos de normalização dos dados são, em geral, não paramétricos como o Lowess e Splines. O gráfico MxA, comumente usado em experimentos com microarrays, é adotado para visualização do padrão de variação das leituras dos alelos A e B e avaliação da necessidade de normalização. A normalização dos dados, em geral, é feita usando as respostas de todos os indivíduos e de todos os SNPs (ou subconjuntos deles) por meio da normalização quantílica, um procedimento não-paramétrico ([Bolstad2003] Bolstad et al., 2003). Modelos paramétricos também podem ser usados na normalização dos dados (ver, por exemplo, [Wolfinger2001] Wolfinger et al., 2001).

Devido à forma que o arranjo é produzido e os experimentos são realizados, uma série de fatores não-biológicos são introduzidos involuntariamente nos dados obtidos. Assim, é necessária a adoção de procedimentos de pré-processamento que visem minimizar (se possível eliminar) os efeitos dos fatores indesejados para que posteriormente, uma análise genômica produza resultados consistentes e confiáveis.

A seguir, vamos descrever os passos do pré-processamento mais utilizados.

E.1 Remoção de Fundo

Os conjuntos de dados provenientes dos chips de SNP são gerados a partir do processamento da imagem escaneada, convertendo-a em medidas quantitativas de luminosidade em cada célula da matriz que forma o arranjo no chip.

No fenômeno da hibridização, o material genético da amostra, anexa-se a seções do microarranjo que possuam sondas de sequências complementares ao alvo genômico. Como o material genético é rotulado com um material fluorescente, regiões onde estas ligações ocorrem com maior frequência aparecem mais iluminadas na imagem escaneada. Regiões do arranjo que não possuem seus complementares presentes no material da amostra devem, teoricamente ter luminosidade nula, porém a interferência da luminescência em regiões vizinhas, adiciona, por menor que seja, um ruído positivo e indeterminado na intensidade daquela seção.

Este fato motiva o seguinte modelo, amplamente empregado em dados de microarranjos, que descreve a intensidade observada, denotada por O , como uma convolução da intensidade de fundo B com o sinal biológico S , conforme a Figura E.1. A remoção de fundo consiste portanto em estimar a intensidade do sinal biológico S , a partir da intensidade observada O .

Figura E.1: O sinal observado em um experimento de *microarray* é composto por um ruído experimental somado ao sinal biológico

Na literatura há várias propostas para este procedimento,

Standard: o método de correção tradicional consiste em usar uma estimativa local do ruído de fundo e subtrair o sinal com esta estimativa, o que pode resultar em valores corrigidos negativos, o que é fisicamente incoerente com o experimento.

Kooperberg: Kooperberg et al. (2002) [?] sugerem o uso do modelo de Bayes Empírico, envolvendo o uso de convoluções ao ruído de fundo e ajustando o sinal de cada spot.

Edwards: uma estratégia simples para evitar valores de intensidades negativos, é proposta por Edwards (2003) [Edwards2003], que ajusta o sinal pela diferença do *background* a partir de um valor mínimo de corte, quando a diferença é menor que o valor de corte, é feito um ajuste por uma função monótona.

NormExp: modelo de convolução normal truncada e exponencial, este é o método utilizado pelo algoritmo RMA descrito em Carvalho et al. (2006) [Carvalho2006]. Os valores corrigidos são estimados por meio da esperança do sinal biológico condicionada aos valores de intensidade observadas. Maiores detalhes são desenvolvidos no Apêndice G.1.

NormExp+offset: uma pequena variação do método NormExp, adicionando um pequeno valor de offset positivo ao valor das intensidades corrigidas para estabilizar a variância.

Vsn: o método de estabilização da variância de Huber et al. (2002) [Huber2002], calibra os dados de cada canal dos *arrays* e usa uma transformação arcoseno hiperbólico generalizada aos dados em vez do logaritmo. A função arcsinh é definida para valores negativos, o que garante que sinais negativos corrigidos possam ser manipulados. Em altas intensidades, a transformação arcoseno hiperbólica é equivalente à logarítmica, enquanto que em baixas intensidades é próximo à diferença entre as intensidades corrigidas.

O trabalho de Matthew et al. (2006)[?] analisa vários métodos de correção disponíveis e sugere por meio de experimentos com dados reais e simulados que o método NormExp+offset possui menor taxa de falsos positivos. Todos os métodos citados acima estão implementados em linguagem R e disponíveis no pacote *limma* desenvolvido por Smyth (2005)[Smyth2005].

E.2 Normalização

Do ponto de vista biológico, a normalização consiste em retirar o viés de caráter experimental dos dados do *microarray*. Do ponto de vista matemático, a normalização consiste em realizar um ajuste, assumindo como premissa, que a maior parte dos SNPs não são diferencialmente expressos.

Considere n medidas observadas (número de *spots*) e, para cada uma delas, duas variáveis aleatórias x_i e y_i , que representam o $\log_2$ das intensidades de cada um dos dois canais alélicos. Os ajustes apresentados a seguir são realizados a partir do gráfico MxA onde $M = x_i - y_i$ e $A = \frac{1}{2}(x_i + y_i)$ (Figura E.2).

Figure E.2: Gráficos MA-plot

Lowess: um procedimento comumente utilizado para a normalização de dados de microarray utiliza uma metodologia de regressão não-paramétrica robusta (denominada LOWESS) para estabilizar a relação entre o logaritmo da razão de intensidades e a média do logaritmo das intensidades em cada lâmina (Yang et al., 2002 [Yang2002a]). Esta metodologia é também utilizada para dados de Affymetrix, mas neste caso o procedimento LOWESS é aplicado sucessivamente para cada par de lâminas, e a normalização final para cada lâmina é dada pela média geral dos resultados de cada um de seus pareamentos. O procedimento é repetido até convergência, i.e., até que pareamentos adicionais não alterem a normalização dos dados, e é denominado LOWESS cíclico. (ver Apêndice H.1)

Splines: em alguns casos, splines cúbicos, podem prover um ajuste superior ao lowess (Bolstad et al., 2003 [Bolstad2003]).

Quantílica: outro procedimento bastante comum de normalização é denominado normalização quantílica (Bolstad et al., 2003 [Bolstad2003]). Este procedimento faz com que todas as amostras apresentem mesma distribuição empírica dos valores de intensidade, de maneira que elas são coincidentes em termos de locação (incluindo medidas de centralidade e percentis) e escala ou variabilidade (ver Apêndice H.2).

E.3 Genotipagem de SNPs em Humanos

Para organismos diplóides, como os humanos, os algoritmos de genotipagem mais comumente usados em dados de SNPs são o Birdseed e o CRLMM (Carvalho et al., 2006 [Carvalho2006]), os quais apresentam um passo de treinamento que usa dados externos do Projeto HapMap para obtenção de estimativas de parâmetros que entrarão no passo de execução dos algoritmos.

Birdseed: fornecido pela Affymetrix para as plataformas 500K, SNP5.0, e SNP6.0, Birdseed usa um algoritmo EM (Expectation-Maximization) para ajustar gaussianas bidimensionais aos dados, produzindo genótipos e índices de confiança para cada indivíduo em cada SNP. Por ser um algoritmo de agrupamento, deve ser utilizado para amostras grandes (tipicamente 50 ou mais).

RLMM: Robust Linear Model with Mahalanobis distance (pronuncia-se *realm*), é um algoritmo de classificação baseado nas técnicas *multi-chip* da Affymetrix. Utiliza uma grande amostra de treinamento com genotipagem conhecida, produzindo melhores resultados que o Birdseed, baseado num modelo linear robusto com distância de Mahalanobis para classificação. A variação não-biológica chip-a-chip é removida através de normalização.

BRLMM: similar ao RLMM, o Bayesian Robust Linear Model with Mahalanobis distance, utiliza informações a priori *ad hoc* a respeito da amostra para encontrar estimativas a posteriori sobre os parâmetros de variabilidade e locação dos agrupamentos.

CRLMM: o Corrected Robust Linear Model with Maximum-likelihood based distances, é um método de classificação supervisionado desenvolvido para SNPs de humanos. Ele utiliza um passo de treinamento com dados do Projeto HapMap para obtenção de parâmetros para o passo de classificação. Carvalho et al. (2006)[Carvalho2006] mostram bons resultados com este algoritmo comparado aos RLMM e BRLMM.

Para o caso poliplóide, todos os algoritmos podem ser estendidos sem grandes dificuldades se a ploidia for conhecida e se dados externos estiverem disponíveis para o passo de treinamento. Porém, na ausência destas informações tem-se um obstáculo adicional: organismos poliplóides podem ter ploidia variável em diferentes SNPs sendo, portanto, necessário o uso de um passo extra para estimação da ploidia dos indivíduos nos correspondentes SNPs e, condicionalmente, realizar o passo de genotipagem. Neste sentido estamos trabalhando no desenvolvimento de um algoritmo em dois estágios, em que no primeiro passo a ploidia dos indivíduos em um SNP é estimada e, no passo seguinte, a classe genotípica é condicionalmente atribuída.

Diferentes abordagens estão sendo consideradas para este problema geral de classificação por meio da definição de funções discriminantes com diferentes custos de classificação errada e diferentes probabilidades a priori. Inicialmente, o caso de mistura de distribuições Normais heterocedásticas está sendo implementado.

Apêndice F

Agrupamentos

A análise de agrupamentos consiste de diversas técnicas e algoritmos que buscam separar (ou agrupar) objetos em grupos semelhantes, tendo como base alguma medida de parença, tal que os elementos pertencentes a um mesmo grupo sejam similares entre si com respeito às características que neles foram medidas, e os elementos em grupos diferentes sejam heterogêneos em relação a estas mesmas características.

As medidas de parença (que são métricas matemáticas) têm um papel central nos algoritmos de agrupamentos, elas são definidas para avaliar se pontos próximos quaisquer podem ou não fazer parte do mesmo agrupamento (ver por ex. Johnson e Wichern, 2002 [JohnsonWichern] , Duda et al., 2001 [?]). Podemos classificá-las em dois tipos:

1. Medidas de Similaridade: quanto maior o valor observado, mais parecidos são os indivíduos (ou objetos).
2. Medidas de Dissimilaridade: quanto maior o seu valor, mais diferentes são os indivíduos (ou objetos).

Geralmente, os algoritmos utilizados na análise de agrupamento estão baseados em medidas de dissimilaridade ou distâncias (métricas).

A idéia básica de agrupamento (clustering) [35, 48] começa quando temos vários objetos e queremos agrupá-los pelas suas propriedades que podem ser quantitativas ou qualitativas. Neste processo existem três parâmetros importantes a serem considerados:

1. a função de distância usada para medir a proximidade (similaridade) entre dois pontos (objetos);
2. a escolha dos pontos que serão usados para medir a distância entre dois grupos;
3. a quantidade de grupos que serão formados.
4. o algoritmo de formação de grupos (hierárquicos ou de aglomeramento).

Uma breve revisão sobre medidas de distância entre agrupamentos, assim como as técnicas mais difundidas na literatura encontram-se no Apêndice ??.

Neste trabalho, não dispomos de uma base de treinamento onde a ploidia e a genotipagem de uma amostra são conhecidas, portanto abordaremos o problema de estimar o número ótimo de agrupamentos no contexto de dados de SNPs.

Apêndice G

Remoção de Fundo

Os modelos descritos neste capítulo são necessários no contexto de plataformas Affymetrix e Illumina, onde é necessário um pré-processamento de dados de expressão para a avaliação de análises estatísticas devido aos erros experimentais amplamente conhecidos nesses experimentos.

G.1 Modelo Normal-Exponencial

Estimação do Sinal Irizarry et al (2003)[Irizarry2003, Irizarry2003] sugere que as distribuições de B e de S sejam respectivamente uma Normal (truncada em zero) e uma Exponencial.

$$O = S + B \quad \begin{cases} S \sim \text{Exp}(\alpha) \\ B \sim NT(\mu, \sigma, 0, \infty) \end{cases}$$

$$\begin{aligned} f_S(s) &= \alpha \exp(-\alpha s) \mathbf{1}_{\{s \geq 0\}} \\ f_B(b) &= \frac{\frac{1}{\sigma} \phi\left(\frac{b-\mu}{\sigma}\right) \mathbf{1}_{\{b \geq 0\}}}{\Phi\left(\frac{\infty-\mu}{\sigma}\right) - \Phi\left(\frac{0-\mu}{\sigma}\right)} = \frac{\frac{1}{\sigma} \phi\left(\frac{b-\mu}{\sigma}\right)}{1 - \Phi\left(-\frac{\mu}{\sigma}\right)} \mathbf{1}_{\{b \geq 0\}} \\ f_O(o) &= \int f_S(o-b) f_B(b) db \end{aligned}$$

$$\begin{aligned} f_O(o) &= \int \alpha \exp(-\alpha(o-b)) \mathbf{1}_{\{o-b \geq 0\}} \frac{\frac{1}{\sigma} \phi\left(\frac{b-\mu}{\sigma}\right)}{1 - \Phi\left(-\frac{\mu}{\sigma}\right)} \mathbf{1}_{\{b \geq 0\}} db \\ &= \int \alpha \exp(-\alpha(o-b)) \frac{\frac{1}{\sigma} \phi\left(\frac{b-\mu}{\sigma}\right)}{1 - \Phi\left(-\frac{\mu}{\sigma}\right)} \mathbf{1}_{\{0 \leq b \leq o\}} db \\ &= \frac{\alpha \exp(-\alpha o)}{1 - \Phi\left(-\frac{\mu}{\sigma}\right)} \int_0^o \exp(\alpha b) \frac{1}{\sigma} \phi\left(\frac{b-\mu}{\sigma}\right) db \\ &= \frac{\alpha \exp(-\alpha o)}{1 - \Phi\left(-\frac{\mu}{\sigma}\right)} \int_0^o \frac{1}{\sigma \sqrt{2\pi}} \exp\left\{\alpha b - \frac{1}{2} \left(\frac{b-\mu}{\sigma}\right)^2\right\} db. \end{aligned}$$

Completando o quadrado em

$$\begin{aligned}
\alpha b - \frac{1}{2} \left(\frac{b-\mu}{\sigma} \right)^2 &= -\frac{1}{2\sigma^2} \left\{ -2\alpha b\sigma^2 + (b-\mu)^2 \right\} = -\frac{1}{2\sigma^2} \left\{ -2\alpha b\sigma^2 + b^2 - 2b\mu + \mu^2 \right\} \\
&= -\frac{1}{2\sigma^2} \left\{ b^2 - 2b(\mu + \alpha\sigma^2) + \mu^2 \right\} \\
&= -\frac{1}{2\sigma^2} \left\{ (b - \mu - \alpha\sigma^2)^2 - 2\mu\alpha\sigma^2 - \alpha^2\sigma^4 \right\} \\
&= -\frac{1}{2\sigma^2} \left\{ (b - \mu - \alpha\sigma^2)^2 - 2\mu\alpha\sigma^2 - \alpha^2\sigma^4 - \mu^2 + \mu^2 \right\} \\
&= -\frac{1}{2} \left\{ \left(\frac{b - \mu - \alpha\sigma^2}{\sigma} \right)^2 - \left(\frac{\mu + \alpha\sigma^2}{\sigma} \right)^2 + \left(\frac{\mu}{\sigma} \right)^2 \right\}. \tag{G.1}
\end{aligned}$$

chegamos a

$$\begin{aligned}
f_O(o) &= \frac{\alpha \exp(-\alpha o)}{(1 - \Phi(-\frac{\mu}{\sigma}))} \frac{\exp\left(-\frac{1}{2} \left(\frac{\mu}{\sigma}\right)^2\right)}{\exp\left(-\frac{1}{2} \left(\frac{\mu + \alpha\sigma^2}{\sigma}\right)^2\right)} \int_0^o \frac{1}{\sigma} \phi\left(\frac{b - \mu - \alpha\sigma^2}{\sigma}\right) db \\
&= \frac{\alpha \exp(-\alpha o)}{(1 - \Phi(-\frac{\mu}{\sigma}))} \frac{\sigma\sqrt{2\pi} \exp\left(-\frac{1}{2} \left(\frac{\mu}{\sigma}\right)^2\right)}{\sigma\sqrt{2\pi} \exp\left(-\frac{1}{2} \left(\frac{\mu + \alpha\sigma^2}{\sigma}\right)^2\right)} \left\{ \Phi\left(\frac{b - \mu - \alpha\sigma^2}{\sigma}\right) \right\}_0^o \\
&= \alpha \exp(-\alpha o) \frac{\phi\left(\frac{\mu}{\sigma}\right)}{\phi\left(\frac{\mu + \alpha\sigma^2}{\sigma}\right)} \left(\frac{\Phi\left(\frac{o - \mu - \alpha\sigma^2}{\sigma}\right) - \Phi\left(\frac{-\mu - \alpha\sigma^2}{\sigma}\right)}{1 - \Phi\left(-\frac{\mu}{\sigma}\right)} \right).
\end{aligned}$$

Vamos calcular agora as condicionais:

$$\begin{aligned}
f_{O|S}(o|s) &= f_B(o - s) = \frac{f_{S,O}(s, o)}{f_S(s)} \\
f_{S|O}(s|o) &= \frac{f_{S,O}(s, o)}{f_O(o)} = \frac{f_{O|S}(o|s) f_S(s)}{f_O(o)} = \frac{f_B(o - s) f_S(s)}{f_O(o)} \\
&= \frac{\frac{1}{\sigma} \phi\left(\frac{o - s - \mu}{\sigma}\right) \alpha \exp(-\alpha s)}{\alpha \exp(-\alpha o) \frac{\phi\left(\frac{\mu}{\sigma}\right)}{\phi\left(\frac{\mu + \alpha\sigma^2}{\sigma}\right)} \left(\frac{\Phi\left(\frac{o - \mu - \alpha\sigma^2}{\sigma}\right) - \Phi\left(\frac{-\mu - \alpha\sigma^2}{\sigma}\right)}{1 - \Phi\left(-\frac{\mu}{\sigma}\right)} \right)} \\
&= \frac{\frac{1}{\sigma} \phi\left(\frac{o - s - \mu}{\sigma}\right) \exp(\alpha(o - s))}{\phi\left(\frac{\mu}{\sigma}\right) \left[\Phi\left(\frac{o - \mu - \alpha\sigma^2}{\sigma}\right) - \Phi\left(\frac{-\mu - \alpha\sigma^2}{\sigma}\right) \right]} \\
&= \frac{\frac{1}{\sigma\sqrt{2\pi}} \exp\left\{ \alpha(o - s) - \frac{1}{2} \left(\frac{o - s - \mu}{\sigma}\right)^2 \right\}}{\phi\left(\frac{\mu}{\sigma}\right) \left[\Phi\left(\frac{o - \mu - \alpha\sigma^2}{\sigma}\right) - \Phi\left(\frac{-\mu - \alpha\sigma^2}{\sigma}\right) \right]}.
\end{aligned}$$

Usando a relação encontrada em (G.1) para $b = o - s$, chegamos a

$$\begin{aligned}
f_{S|O}(s|o) &= \frac{\phi\left(\frac{\mu+\alpha\sigma^2}{\sigma}\right)}{\phi\left(\frac{\mu}{\sigma}\right)\left[\Phi\left(\frac{o-\mu-\alpha\sigma^2}{\sigma}\right) - \Phi\left(\frac{-\mu-\alpha\sigma^2}{\sigma}\right)\right]} \frac{\phi\left(\frac{\mu}{\sigma}\right)}{\phi\left(\frac{\mu+\alpha\sigma^2}{\sigma}\right)} \frac{1}{\sigma} \phi\left(\frac{o-s-\mu-\alpha\sigma^2}{\sigma}\right) \\
&= \frac{\mathbf{1}_{\{0 \leq s \leq o\}}}{\left[\Phi\left(\frac{o-\mu-\alpha\sigma^2}{\sigma}\right) - \Phi\left(\frac{-\mu-\alpha\sigma^2}{\sigma}\right)\right]} \frac{1}{\sigma} \phi\left(\frac{o-s-\mu-\alpha\sigma^2}{\sigma}\right) \\
&= \frac{\mathbf{1}_{\{0 \leq s \leq o\}}}{\left[\Phi\left(\frac{o-\mu-\alpha\sigma^2}{\sigma}\right) - \Phi\left(\frac{-\mu-\alpha\sigma^2}{\sigma}\right)\right]} \frac{1}{\sigma} \phi\left(\frac{s-(o-\mu-\alpha\sigma^2)}{\sigma}\right) \\
&= NT(o-\mu-\alpha\sigma^2, \sigma^2, 0).
\end{aligned}$$

$$\begin{aligned}
E[S|O=o] &= \int s f_{S|O}(s|o) ds \\
&= \int_0^o \frac{s}{\left[\Phi\left(\frac{o-\mu-\alpha\sigma^2}{\sigma}\right) - \Phi\left(\frac{-\mu-\alpha\sigma^2}{\sigma}\right)\right]} \frac{1}{\sigma} \phi\left(\frac{o-s-\mu-\alpha\sigma^2}{\sigma}\right) ds \\
&= - \frac{\int_0^o \frac{(-s)}{\sigma} \phi\left(\frac{o-s-\mu-\alpha\sigma^2}{\sigma}\right) ds}{\left[\Phi\left(\frac{o-\mu-\alpha\sigma^2}{\sigma}\right) - \Phi\left(\frac{-\mu-\alpha\sigma^2}{\sigma}\right)\right]}. \tag{G.2}
\end{aligned}$$

Integrando $\int_a^b x \phi\left(\frac{x-m}{s}\right) dx$, fazendo $x = \sigma y + \mu$, temos

$$\begin{aligned}
\int_a^b x \phi\left(\frac{x-m}{s}\right) dx &= \int_{\frac{a-\mu}{\sigma}}^{\frac{b-\mu}{\sigma}} (sy + m) \phi(y) \sigma dy \\
&= s \int_{\frac{a-m}{s}}^{\frac{b-m}{s}} y \phi(y) dy + m \int_{\frac{a-m}{s}}^{\frac{b-m}{s}} \phi(y) dy \\
&= s \int_{\frac{a-m}{s}}^{\frac{b-m}{s}} y \phi(y) dy + m \left(\Phi\left(\frac{b-m}{s}\right) - \Phi\left(\frac{a-m}{s}\right) \right) \\
&= s \left(\phi\left(\frac{a-m}{s}\right) - \phi\left(\frac{b-m}{s}\right) \right) + m \left(\Phi\left(\frac{b-m}{s}\right) - \Phi\left(\frac{a-m}{s}\right) \right)
\end{aligned}$$

Fazendo $a = 0$, $b = o$, $x = -s$ e $m = -o + \mu + \alpha\sigma^2$ e $s = \sigma$ em (G.2), vemos

$$\begin{aligned}
E[S|O=o] &= \frac{-\sigma \left(\phi\left(\frac{-\mu-\alpha\sigma^2}{\sigma}\right) - \phi\left(\frac{o-\mu-\alpha\sigma^2}{\sigma}\right) \right) - (-o + \mu + \alpha\sigma^2) \left[\Phi\left(\frac{o-\mu-\alpha\sigma^2}{\sigma}\right) - \Phi\left(\frac{-\mu-\alpha\sigma^2}{\sigma}\right) \right]}{\left[\Phi\left(\frac{o-\mu-\alpha\sigma^2}{\sigma}\right) - \Phi\left(\frac{-\mu-\alpha\sigma^2}{\sigma}\right) \right]} \\
&= (o - \mu - \alpha\sigma^2) - \sigma \left[\frac{\phi\left(\frac{-\mu-\alpha\sigma^2}{\sigma}\right) - \phi\left(\frac{o-\mu-\alpha\sigma^2}{\sigma}\right)}{\Phi\left(\frac{o-\mu-\alpha\sigma^2}{\sigma}\right) - \Phi\left(\frac{-\mu-\alpha\sigma^2}{\sigma}\right)} \right]
\end{aligned}$$

Fazendo $a = o - \mu - \alpha\sigma^2$ e $b = \sigma$

$$\begin{aligned}
E[S|O=o] &= a - b \left[\frac{\phi\left(\frac{a-o}{b}\right) - \phi\left(\frac{a}{b}\right)}{\Phi\left(\frac{a-o}{b}\right) - \Phi\left(\frac{a}{b}\right)} \right] \\
&= a + b \left[\frac{\phi\left(\frac{a}{b}\right) - \phi\left(\frac{o-a}{b}\right)}{\Phi\left(\frac{a}{b}\right) - \Phi\left(\frac{a-o}{b}\right)} \right]
\end{aligned}$$

Lembrando que $\Phi(x) = 1 - \Phi(-x)$, temos então que:

$$E[S|O=o] = a + b \left[\frac{\phi\left(\frac{a}{b}\right) - \phi\left(\frac{o-a}{b}\right)}{\Phi\left(\frac{a}{b}\right) + \Phi\left(\frac{o-a}{b}\right) - 1} \right]$$

G.2 Estimação dos parâmetros do modelo Normal-Exponencial:

Apesar de ser possível a aplicação de métodos de Newton ou algoritmos EM para obtermos estimativas de máxima verossimilhança para o modelo de convolução descrito no Apêndice anterior, a instabilidade e a alta dependência com relação ao ponto inicial, torna o uso destes métodos ineficazes, como mostra McGee e Chen (2006) [?], além disso, o volume de dados torna o uso do algoritmo EM impraticável pela complexidade da verossimilhança.

Para contornar este problema McGee e Chen (2006)[McGee2006] observam que é possível apresentar estimativas por métodos ad hoc compatíveis com a natureza do modelo.

Encontrando a moda de O , obtemos:

$$\frac{\partial l}{\partial o} = -\alpha + \frac{\frac{1}{\sigma}\phi\left(\frac{o-\mu-\alpha\sigma^2}{\sigma}\right)}{\Phi\left(\frac{o-\mu-\alpha\sigma^2}{\sigma}\right) - \Phi\left(\frac{-\mu-\alpha\sigma^2}{\sigma}\right)}$$

Se o_m for a moda do sinal observado, zerando a equação anterior, chegamos à relação:

$$\phi\left(\frac{o_m - \mu}{\sigma} - \alpha\sigma\right) = \alpha\sigma \left[\Phi\left(\frac{o_m - \mu}{\sigma} - \alpha\sigma\right) - \Phi\left(-\frac{\mu}{\sigma} - \alpha\sigma\right) \right] \quad (G.3)$$

Apesar das simplificações feitas para encontrar as derivadas da verossimilhança, os métodos de Newton tradicionais são instáveis e dependentes do ponto inicial. Uma alternativa é utilizar métodos *ad hoc* que façam sentido no contexto do modelo, como o usado no algoritmo RMA.

Considere $m = \max f_O(x)$, o valor de x onde f_O atinge seu máximo vamos chamá-lo de x_m .

M1

- $\hat{\mu} = \arg \max [f(x|x < x_m)]$, ou seja, dado x_m , a estimativa da média é a moda das intensidades abaixo de x_m .
- $\hat{\sigma} = \sqrt{2} \cdot SD(x|x < \hat{\mu})$ é o desvio amostral das intensidades observadas que são menores que $\hat{\mu}$, multiplicado por um fator $\sqrt{2}$.
- $\hat{\alpha} = [\arg \max [f(x|x > x_m)] - x_m]^{-1}$

M2 As estimativas de $\hat{\mu}$ e $\hat{\sigma}$ são as mesmas do método 1, para estimar $\hat{\alpha}$, utiliza-se a média, a mediana e os percentis 75 e 99,99 dos dados maiores que a moda geral.

M3 $\hat{\mu} = (\mu_1 + \mu_2)/2$, onde μ_1 , σ_1 e α_1 são estimados pelo método 1, μ_2 é obtido usando-se a relação (G.3). Uma vez calculada, $\hat{\mu}$ é utilizada para estimar $\hat{\sigma}$ e $\hat{\alpha}$ como no M1.

M4 Em vez de estimar $\hat{\mu}$ usando a moda dos dados menores que x_m , que tende a superestimar a verdadeira média, $\hat{\mu}$ é encontrada usando-se a moda das intensidades menores que $2x_m$, $\hat{\sigma}$ é calculada como no método 1 e $\hat{\alpha}$ é dado por vários percentis de observações maiores que a média estimada (M2).

M5 $\hat{\mu}$ e $\hat{\sigma}$ são estimados como em M4 e $\hat{\alpha}$ em M2. Valores corrigidos dos parâmetros são encontrados usando-se a relação (G.3). Tira-se a média entre as estimativas corrigidas com as originais.

M6 $\hat{\mu} = x_m$, $\hat{\sigma} = SD(x|x < \hat{\mu})$, α é estimado usando-se a média e os percentis de todas as intensidades observadas (e não apenas as maiores que x_m).

M7 $\hat{\mu}$ é a média entre x_m e a estimativa dada pela correção (G.3), σ e α são estimados usando M6.

Apêndice H

Métodos de Normalização

H.1 Lowess

LOWESS (LOcally WEighted polynomial regreSSion), originalmente proposto por Cleveland (1979) [Cleveland79] e desenvolvido pelo Cleveland e Devlin (1988)[Cleveland88], é um método de regressão polinomial localmente ponderada. Em cada ponto do conjunto de n dados um polinômio de baixo grau é ajustado a um subconjunto de dados, com valores das variáveis preditoras perto do ponto de cuja resposta está sendo estimada. O polinômio é ajustado usando mínimos quadrados ponderados, dando mais peso aos pontos perto do ponto de cuja resposta está sendo estimado e menos peso aos pontos mais distantes. O valor da função de regressão para o ponto é, então, obtido através da avaliação polinomial local. O ajuste Lowess termina após os valores da função de regressão forem calculadas para cada um dos pontos do conjunto de dados.

Os subconjuntos utilizados para cada regressão de mínimos quadrados ponderados ajustados no Loess são determinados por um algoritmo de vizinhos mais próximos. O usuário deve determinar uma "banda" ou "parâmetro de suavização", que define a quantidade de dados que é usada para ajustar cada polinômio local. O parâmetro de alisamento, α , é um número entre $\frac{\lambda+1}{n}$ e 1, com λ denota o grau do polinômio local. O valor de α é a proporção de dados utilizados em cada ajuste. O subconjunto utilizado em cada ajuste dos mínimos quadrados ponderados compreende a $n\alpha$ (arredondado para o próximo inteiro maior) pontos preditores cujos valores estão mais próximos do ponto em que a resposta está sendo estimada.

Os polinômios ajustados em cada subconjunto de dados são quase sempre de primeiro ou segundo graus, ou seja, localmente lineares ou quadráticos. O uso de polinômios de grau zero torna o Lowess uma média ponderada móvel. Esse modelo simples pode funcionar bem em algumas situações, mas nem sempre. Polinômios de alto grau podem produzir modelos instáveis por conta do efeito de overfitting aos dados. Lowess é baseado na idéia de que qualquer função pode ser bem aproximada em uma pequena vizinhança por um polinômio de ordem baixa e que modelos simples podem se adequar aos dados facilmente.

Como mencionado acima, a função de peso dá mais peso aos pontos de dados mais próximos do ponto de estimação e menos peso aos pontos de dados que estão mais afastados. A utilização dos pesos baseia-se na idéia de que pontos próximos uns dos outros no espaço preditor são mais prováveis de serem correlacionados uns aos outros do que os pontos mais afastados. Seguindo esta lógica, os pontos que ajustam-se melhor ao modelo localmente possuem maior influência na estimação de parâmetros locais, pontos mais afastados, possuem uma influência menor.

Uma função peso tradicional no Lowess é a função tri-cúbica

$$W(u) = \begin{cases} (1 - |u|^3)^3 & \text{para } |u| < 1 \\ 0 & \text{para } |u| \geq 1 \end{cases}$$

No entanto, qualquer outra função do peso que satisfaça as propriedades listadas em Cleveland (1979)[Cleveland79] também pode ser usada. O peso de um ponto específico de qualquer subconjunto local é obtido através da avaliação da W para a distância entre esse ponto e o ponto de estimação.

Figura H.1: Antes e depois da normalização Lowess

H.2 Normalização quantílica

Segundo Bolstad et al (2003) [Bolstad2003] objetivo da normalização quantílica é impor a mesma distribuição empírica das intensidades para cada array. O gráfico quantil-quantil irá ter uma linha diagonal reta com inclinação de 45 graus se dois vetores de dados tem a mesma distribuição. Assim, plotando os quantis de dois vetores de dados um contra o outro e então projetando cada ponto sobre uma linha diagonal de 45 graus isso irá produzir uma transformação que proporciona a mesma distribuição para ambos os vetores de dados.

Em n dimensões, um gráfico quantil-quantil plot onde todos os vetores de dados possuem a mesma distribuição teriam os pontos sobre a linha descrita pelo vetor $\left(\frac{1}{\sqrt{n}}, \frac{1}{\sqrt{n}}, \dots, \frac{1}{\sqrt{n}}\right)$. Essa extensão para n dimensões motivou o algoritmo da normalização quantílica.

Algoritmo da normalização quantílica:

1. Sejam n vetores de comprimento p , de X , de dimensão $p \times n$, onde cada array é uma coluna de X .
2. Organize em ordem crescente cada coluna de X e chame de X_s .
3. Calcule as médias das linhas de X_s e crie X'_s , um arranjo da mesma dimensão de X , mas onde todos os valores de cada linha são iguais as médias da respectiva linha em X_s .
4. Calcule X_n reorganizando cada coluna de X'_s para ter a mesma ordem do correspondente vetor de entrada X .

Esta é a normalização padrão do algoritmo RMA (Robust Multichip Average) descrito por . Observe na Figura 26 que após a normalização quantílica as quatro hibridizações passaram a ter a mesma distribuição e os seus gráficos ficaram sobrepostos. Pode-se observar na Figura H.2 que as hibridizações tornaram-se comparáveis observando-se seus boxplots antes e depois da normalização quantílica.

(a) Densidades

(b) Boxplots

Figura H.2: Antes e depois da normalização quantílica

Apêndice I

Técnicas de Agrupamento

I.1 Medidas de Distância

A distância entre dois pontos mede a similaridade que existe entre dois objetos. Sejam $x = (x_1, \dots, x_p)$ e $y = (y_1, \dots, y_p)$ duas observações (p-dimensionais) e $d(x, y)$ a distância entre elas. A seguir são apresentadas algumas medidas mais utilizadas.

Euclidiana $d(x, y) = \sqrt{\sum_{i=1}^p (x_i - y_i)^2}$ (esta é a mais popular).

Minkowski $d(x, y) = \sqrt[m]{\sum_{i=1}^p |x_i - y_i|^m}$, com $m = 1$, temos a Manhattan, $m = 2$ temos a Euclidiana, variando o m mudamos os pesos atribuídos a pequenas e grandes diferenças.

Camberra $d(x, y) = \sum_{i=1}^p \frac{|x_i - y_i|}{(x_i - y_i)}$, apenas para valores não-negativos.

Czkanowski $d(x, y) = 1 - 2 \frac{\sum_{i=1}^p \min_{1 \leq i \leq p} (x_i, y_i)}{\sum_{i=1}^p (x_i - y_i)}$, apenas para valores não-negativos.

Máximo $d(x, y) = \max_{1 \leq i \leq p} \{|x_i - y_i|\}$.

Mahalanobis $d(x, y) = \sqrt{(x - y)' \Sigma^{-1} (x - y)}$, onde Σ é a matriz de covariância entre x e y .

I.2 Métodos de União

Para avaliar a similaridade entre dois grupos, ou um grupo e um elemento, precisamos obter um ponto que represente-o e seja usado no cálculo de $d(x_i, y_i)$. A seguir apresentamos alguns métodos.

Vizinho mais próximo (Single) Escolhe de cada grupo um ponto de tal forma que $d(x_i, y_i)$ seja a menor possível.

Vizinho mais distante (Complete) Escolhe de cada grupo um ponto de tal forma que $d(x_i, y_i)$ seja a maior possível.

Centróide (Average) $d(r, s) = d(\bar{x}_r, \bar{x}_s)$ onde r e s são dois grupos e, $\bar{x}_r$ e $\bar{x}_s$ seus respectivos centróides ou pontos médios.

Divisão (Ward) $d^*(r, s) = \frac{n_r n_s}{n_r + n_s} d(\bar{x}_r, \bar{x}_s)$, onde n_r e n_s são o número de elementos de cada grupo.

Correlação $d(x, y) \equiv \rho(x, y) \in (-1, 1)$. Tanto na análise de agrupamentos quando na de classificadores, a medida de correlação é usada com bons resultados (Dudoit et al. (2002)[Dudoit2002-1]). Quanto maior a correlação, menor a distância entre os grupos. Esta medida é equivalente à Euclidiana se aplicada nos dados padronizados.

I.3 Agrupamento Hierárquico

Neste método de agrupamento a quantidade de grupos a serem formados não é um parâmetro, pelo fato de que o resultado final é um grupo com todos os objetos. A idéia consiste em colocar as uniões numa hierarquia decrescente de similaridade. Primeiro são agrupados dois objetos com o a maior similaridade, depois este grupo de dois objetos é unido ao terceiro objeto mais similar e assim sucessivamente.

I.4 Agrupamento k-médias

Neste divide-se os objetos em k partições. A partir do método de união escolhido, para cada grupo é determinado seu ponto de referência, e para cada objeto em particular é determinada a sua similaridade com cada um dos grupos, usando os pontos de referência e a função de distância definida. Se algum dos objetos for mais similar com outro grupo que não seja o dele, então este é realocado. O algoritmo é iterado, recalculando os pontos de referência, procurando aqueles que ainda não estejam no grupo mais apropriado e realocando os objetos novamente. Um objeto é realocado a cada iteração, mas apenas dois pontos de referência precisam ser recalculados (os que envolvem os dois grupos modificados). O processo de realocação termina com o posicionamento de cada objeto no grupo onde ele melhor se enquadra (ou seja, naquele grupo de objetos mais similares a ele), ou quando o algoritmo atinge um número máximo de iterações indicado pelo usuário.

O resultado final vai depender dos parâmetros fornecidos, número de grupos, agrupamentos iniciais (definidos aleatoriamente ou fornecidos pelo usuário), função de distância, método de união e número de iterações. Pode ocorrer casos em que o agrupamento final não estabilize mesmo com número de iterações suficientemente grande. Também é possível que mesmo fixados os parâmetros iniciais, várias execuções resultem em agrupamentos não idênticos.

Índice Remissivo

D

DNA, 1

N

nucleotídeo, 1

P

ploidia, 1

S

SNP, 1